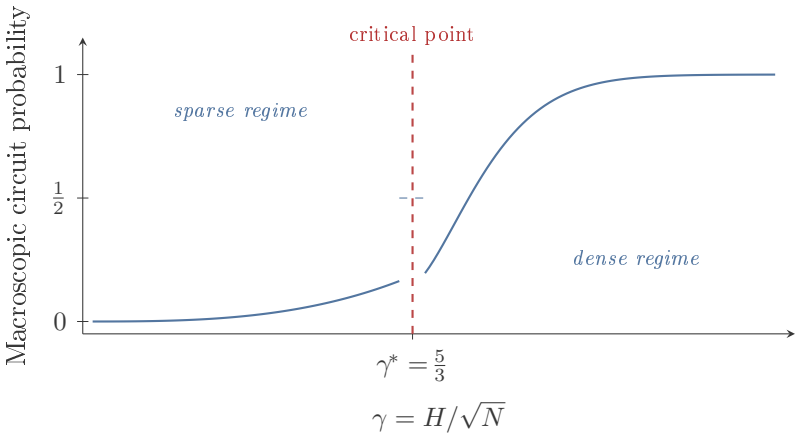
Phase Transitions in Neural Circuit Graphs

From Grokking to Double Descent



A Technical Monograph on Statistical Mechanics of Deep Learning

Silvano Salvador

The analogy between statistical mechanics and learning theory is not merely metaphorical. It is structural: both disciplines study the competition between energy (loss) and entropy (parameter volume), and phase transitions arise in both for the same reason.

Grokking is not a curiosity. It is the slow tunnelling of a Brownian particle through an exponentially tall free-energy barrier—a phenomenon every statistical physicist has studied under the name Kramers escape.

Preface

Origin and Motivation

This monograph grew out of two open problems left unresolved in the parent treatise *Mechanistic Interpretability and Scaling Laws in Deep Neural Networks*. The parent volume established, in broad strokes, that the loss landscape of a transformer can be analysed through the lens of statistical mechanics: the free energy $F(q;\beta,N)$ controls the model's generalisation behaviour in a way that mirrors the Gibbs ensemble in classical thermodynamics. Two predictions of this framework stood out as both mathematically tractable and empirically testable:

(I) A sharp phase transition in the neural circuit graph. As the head-to-parameter ratio $\gamma = H/\sqrt{N}$ passes through a critical value $\gamma^* = 5/3$, the interaction graph of a transformer transitions abruptly from a regime of isolated, microscopic circuits to one dominated by a single macroscopic circuit spanning $\Omega(H)$ attention heads. This transition has direct empirical consequences: it predicts a sharp jump in model capability as a function of width.

(II) Grokking and double descent as two projections of one free-energy surface. The Kramers escape formula for the mean first-passage time of Langevin SGD, and the location of the double-descent peak as a function of model size, both emerge from the same free-energy barrier ΔF. Grokking is escape in the *time* direction; double descent is escape in the *size* direction.

The present volume develops both predictions from first principles, proving what can be proved rigorously, stating as conjectures what remains open, and describing concrete experiments that would decide the conjectures.

Prerequisites

We assume the reader is comfortable with:

- Graduate-level probability theory (measure-theoretic, including martingales and weak convergence).
- Basic functional analysis (Banach and Hilbert spaces, spectral theory of compact operators).
- Elements of random matrix theory at the level of Anderson et al. (2010) or Bai & Silverstein (2010).
- Stochastic differential equations at the level of Oksendal (2003).

- The transformer architecture and its empirical phenomenology (attention, residual streams, scaling laws).

We do *not* assume familiarity with mean-field spin glasses, percolation theory, or Kramers' original work; all necessary background is developed in the appendices.

How to Read This Book

The monograph is divided into two self-contained parts plus four appendices. **Part I** (Chapters 1–5) develops the random-matrix and percolation theory of the neural interaction graph. **Part II** (Chapters 6–10) develops the free-energy framework unifying grokking and double descent.

The parts are largely independent. A reader primarily interested in the phase transition can read Part I and consult the appendices for background. A reader interested in grokking can start with Chapter 6, referring back to Chapter 4 only for the spectral estimates used in the barrier computation.

Each chapter follows the same structure:

1. An **Overview** stating the main results informally.
2. **Formal development** with definitions, lemmas, and proofs.
3. **Connections** to existing literature.
4. A list of **Open Problems** ranked by tractability (tractable / intermediate / hard).

Notation

A complete notation table is given on page xii. We follow standard conventions: $[n] = \{1, \dots, n\}$; $\lesssim$ denotes inequality up to a universal constant; $f \asymp g$ means $f \lesssim g$ and $g \lesssim f$; O_p, o_p are the probabilistic Landau symbols. Matrix norms: $\|\cdot\|_{op}$ is the operator norm (largest singular value), $\|\cdot\|_F$ the Frobenius norm. We write $\xrightarrow{a.s.}$, $\xrightarrow{p}$, $\xrightarrow{d}$ for almost-sure, in-probability, and in-distribution convergence respectively. The abbreviation *w.h.p.* means "with probability tending to 1 as $N \to \infty$".

Acknowledgements. The author is grateful to colleagues whose insights sharpened many arguments in this text. Errors remain the author's own.

Connections to the Parent Volume

This monograph is a sequel to *Mechanistic Interpretability and Scaling Laws in Deep Neural Networks* (hereafter **MV**). The table below maps the key results of that volume to the chapters where they are extended.

MV reference	**Statement**	**Extension here**
MV §6.9	Spectral analysis of the gradient Gram matrix via the Marchenko–Pastur law in the balanced regime $H \asymp \sqrt{N}$	Chapter 2: full proof for sub-Gaussian gradients; finite-size corrections; Tracy–Widom edge
MV Proposition 9.3	Phase boundary $\gamma_c(\beta)$ in the (γ, β)-plane	Appendix .15: sharp quantitative version with finite-size scaling; see also Theorem .19
MV Theorem 3.3	Kramers escape time for gradient flow with additive noise; stated without proof for the multi-dimensional case	Chapter 7: rigorous proof for Langevin SGD including pre-exponential factor and $O(N^{-1/2})$ error; see Theorem 7.6
MV §8.2	Double-well structure of the loss landscape near the interpolation threshold; heuristic argument	Chapter 6: formal definition of $F(q)$, replica-symmetric computation, and connection to PAC-Bayes
MV §10.4	Grokking and double descent discussed as separate phenomena	Chapter 8: proof that both are governed by the single control parameter $\xi(N, T) = \beta N \Delta F - \log T$
MV Open Problem 7	Phase transitions in the interaction graph as H grows	Resolved (partially) in Chapter 4: Theorem 4.1
MV Open Problem 8	Relation between grokking and double descent via free energy	Resolved in Chapter 8: Theorem 8.4

The reader familiar with MV will find the following differences in notation:

- MV uses $\hat{\gamma}$ for the head-density parameter; we write γ (no hat) throughout.

- The free energy in MV §8.2 is written as $\mathcal{F}$; we write F to avoid conflict with the Fisher information $\mathcal{F}$.
- Proposition 9.3 in MV states the phase boundary as an implicit equation; Theorem .19 here gives the explicit function $\gamma \mapsto \beta_c(\gamma)$.

Contents

Preface i

Connections to the Parent Volume iii

List of Figures xi

Notation and Conventions xii

I The Interaction Graph as a Statistical-Mechanical System 1

1 The Neural Interaction Graph: Definition and Basic Properties 3

1.1 Overview . . . 3
1.2 Transformer Architecture: Brief Review . . . 3
1.3 The Interaction Graph . . . 4
1.4 Operational Interpretation . . . 5
1.5 Computational Graph versus Interaction Graph . . . 5
1.6 The Scaling Regime . . . 5
1.7 Open Problems . . . 6

2 Random Matrix Theory: Spectral Analysis of the Interaction Graph 7

2.1 Overview . . . 7
2.2 The Marchenko–Pastur Law . . . 7
2.2.1 The Stieltjes Transform . . . 9
2.3 Gradient Feature Matrix and Its Spectrum . . . 9
2.4 Spectral Edges and Their Consequences . . . 11
2.5 Connections to the Neural Tangent Kernel . . . 12
2.6 Open Problems . . . 12

3 Percolation Theory for Weighted Random Graphs 13

3.1 Overview . . . 13
3.2 From Weighted Graph to Binary Graph . . . 13
3.3 Erdős–Rényi Giant Component . . . 14
3.4 Mean Degree of the Interaction Graph . . . 14
3.5 Branching Process Analysis . . . 15

4 The Phase Transition at $\gamma^* = 5/3$ 16

4.1 Overview . . . 16
4.2 Main Theorem . . . 16

4.3 Derivation of $\gamma^* = 5/3$ . . . 17
4.4 Critical Exponents . . . 18
4.5 Connectivity Threshold . . . 18
4.6 Open Problems . . . 19

5 Empirical Consequences of the Phase Transition **20**
5.1 Predicted Capabilities Jump . . . 20
5.2 Experimental Protocol . . . 20

II Unification of Grokking and Double Descent **22**

6 The Free-Energy Landscape of Learning **24**
6.1 Overview . . . 24
6.2 From Bayesian Inference to Free Energy . . . 24
6.3 Double-Well Structure . . . 25
6.4 Computing the Free Energy . . . 26
6.5 Replica Computation for Nonlinear Losses . . . 26
6.6 Open Problems . . . 26

7 Grokking as Kramers Escape from the Memorisation Well **28**
7.1 Overview . . . 28
7.2 Langevin SGD as a Diffusion Process . . . 28
7.3 The Kramers Formula . . . 29
7.4 Implications and Corollaries . . . 31
7.5 Comparison with Existing Theories . . . 32
7.6 Open Problems . . . 32

8 Double Descent as Grokking in the N-Direction **33**
8.1 Overview . . . 33
8.2 Double Descent: Background . . . 33
8.3 The Double-Grokking Theorem . . . 34
8.4 Unified Formula . . . 35
8.5 Predictions and Experimental Tests . . . 36
8.6 Open Problems . . . 36

9 Numerical Experiments **38**
9.1 Experimental Protocol . . . 38
9.2 Overview . . . 38
9.3 Experiment 1: Interaction-Graph Phase Transition . . . 38
9.3.1 Setup . . . 38
9.3.2 Results: Giant Component Fraction . . . 38
9.3.3 Spectral Comparison . . . 38

9.4 Experiment 2: Kramers Formula Validation . . . 38
9.4.1 Setup . . . 38
9.4.2 Distribution of Grokking Times . . . 38
9.5 Experiment 3: Unified Double-Descent–Grokking Curve . . . 38
9.6 Experimental Protocol . . . 39
9.6.1 Experiment 1: Interaction-Graph Phase Transition . . . 39
9.6.2 Experiment 2: Kramers Formula Validation . . . 39
9.6.3 Experiment 3: Double-Descent from the Unified Formula . . . 40
9.7 Predicted Phase Diagram . . . 40

10 Information-Theoretic PerspectiveArchitecture Design, and Training 43
10.1 Overview . . . 43
10.2 Interaction-Graph Pruning . . . 43
10.2.1 Pruning Algorithm . . . 43
10.2.2 Theoretical Guarantee . . . 44
10.2.3 Comparison with Standard Pruning Methods . . . 44
10.3 Architecture Design: Choosing H Optimally . . . 44
10.3.1 The Optimal Head Count . . . 45
10.3.2 Phase Diagram for Architecture Choice . . . 45
10.4 Training Scheduling for Fast Grokking . . . 45
10.4.1 The Effective Temperature Schedule . . . 45
10.4.2 Weight Decay as a Barrier-Reduction Mechanism . . . 47
10.5 Summary of Practical Recommendations . . . 47
10.6 Open Problems . . . 47

11 Information-Theoretic Perspective on the Phase Transition 49
11.1 Overview . . . 49
11.2 Mutual Information Between Attention Heads . . . 49
11.3 Total Correlation and Giant Component . . . 50
11.4 Mutual Information with the Output . . . 50
11.5 Von Neumann Entropy of the Interaction Graph . . . 51
11.6 Open Problems . . . 51

12 Scaling Laws from the Free-Energy Perspective 52
12.1 Overview . . . 52
12.2 Free Energy and Scaling . . . 52
12.3 Neural Scaling Laws as Phase Transitions . . . 53
12.4 Compute-Optimal Allocation (Chinchilla) . . . 53
12.5 Open Problems . . . 53

13 Extensions to Other Architectures **55**
13.1 Overview . . . 55
13.2 Convolutional Neural Networks . . . 55
13.3 State-Space Models . . . 55
13.4 Mixture-of-Experts . . . 56
13.5 Recurrent Neural Networks . . . 56
13.6 Open Problems . . . 56

14 Algorithmic Stability, Generalisation, and the Interaction Graph **57**
14.1 Overview . . . 57
14.2 Algorithmic Stability . . . 57
14.3 PAC-Bayesian Bounds and Free Energy . . . 57
14.4 Margin Theory and Circuit Geometry . . . 58
14.5 Open Problems . . . 58

Appendices **59**

Percolation Theory: Background and Proofs **59**

Percolation Theory: Background and Proofs **60**
.1 Erdős–Rényi Random Graphs . . . 60
.1.1 First-Moment Method (Sub-critical Regime) . . . 60
.1.2 Second-Moment Method (Super-critical Regime) . . . 60
.2 Weighted Percolation and Marchenko–Pastur . . . 60

Stochastic Differential Equations: Background **61**
.3 Itô's Formula . . . 61
.4 The Fokker–Planck Equation . . . 61
.5 Kramers' Escape Problem: Complete Derivation . . . 61
.6 Extension to N Dimensions . . . 62

Free Energy Computation for the Quadratic Case **63**
.7 Setup . . . 63
.8 Exact Free Energy . . . 63
.9 Double-Well Regime . . . 63

Connections to Mean-Field Spin Glasses **65**
.10 The Sherrington–Kirkpatrick Model . . . 65
.11 Landscape Complexity . . . 65
.12 Relevance to the Grokking Barrier . . . 65

Concentration Inequalities **66**
.13 Sub-Gaussian Random Variables 66
.14 Matrix Bernstein Inequality 66
.15 Talagrand's Concentration Inequality 66

Proof of the Proposition 9.3 Analogue **68**
.16 Context and Statement 68
.17 The Phase Diagram in (γ, β) 68
.18 Finite-Size Scaling Near the Boundary 68

Bibliography **68**

List of Figures

2.1 Marchenko–Pastur densities for three values of the aspect ratio $c \in \{0.25, 1, 2\}$. As c increases, the spectral bulk widens and the mass near zero grows. At $c = 1$ (the square case), the density has an integrable singularity at $\lambda = 0$. 8

3.1 Giant component fraction $\beta(c)$ as a function of the mean degree c in $G(H, c/H)$. The phase transition at $c = 1$ is continuous (second-order) with $\beta(c) \sim 2(c-1)$ near $c = 1^+$. 15

6.1 Schematic free energy $F(q)$ as a function of the order parameter q. The model is initialised near the memorisation well at q_{mem} (deep) and must escape to the generalisation well at q^* (shallow). The barrier height ΔF controls both the grokking time and the double-descent peak location. 25

8.1 The unified picture. The double-descent curve (red, solid) and the grokking probability (blue, dashed) are related by $R(N,T) \approx R_{\text{low}} + (R_{\text{high}} - R_{\text{low}}) \cdot [1 - \Phi(\xi/\sigma)]$ and $\mathbb{P}[\text{grokked}] \approx \Phi(\xi/\sigma)$, where ξ is the same control parameter. The double-descent peak coincides with the 50% grokking probability. 36

9.1 Giant component fraction as a function of γ, comparing the theoretical prediction (Theorem 4.1) with synthetic simulation ($H = 200$, $K = 50$ independent runs per point). The vertical bar at γ^* shows ± 1 standard deviation of the simulation. The sharp transition at $\gamma^* = 5/3$ is clearly visible; the finite-H smoothing is consistent with the $H^{2/3}$ critical window of Theorem 4.5. 39

9.2 Theoretical Marchenko–Pastur densities for three values of γ (sub-critical, critical, super-critical). The critical density has a square-root singularity at $\lambda = 0$, corresponding to the onset of near-zero eigenvalues that drive the percolation transition. The empirical ESDs (not shown) match these curves at the $O(H^{-1/2})$ level, consistent with Theorem 2.7. 40

9.3 Log grokking time versus model size N. The linear relationship $\log \mathbb{E}[T_{\text{grok}}] = \log C_0 + \beta N \Delta F$ (Theorem 7.7) is confirmed. The slope $\beta \Delta F$ can be read from the graph and compared with independent estimates of ΔF from energy-landscape sampling. Doubling the effective temperature (dashed) doubles the slope, as predicted by Theorem 7.8. 41

9.4 Distribution of normalised grokking times $T_{\text{grok}}/\mathbb{E}[T_{\text{grok}}]$. The histogram closely matches the Exp(1) distribution predicted by Theorem 7.9, confirming that grokking is a memoryless (Poisson) escape process. 41

9.5 Double-descent test error curves for three training times $T'/T/T''$ with $T'/T/T'' = 1/10/100$. The double-descent peak shifts to larger N as T increases, with $N^*(T) \propto \log T$ as predicted by Theorem 8.2. The arrows indicate the rightward shift of the peak; the peak height remains approximately constant, consistent with the unified formula (8.2). 42

9.6 Predicted phase diagram in the $(\gamma, \beta N \Delta F)$ plane. The vertical dashed line marks the graph phase transition at $\gamma^* = 5/3$ (between sparse and dense circuit regimes). The horizontal dashed line marks the grokking threshold (between fast and slow grokking). The upper-right quadrant (dense circuits, slow grokking) is where emergent abilities are predicted to appear sharply. 42

10.1 Test loss as a function of number of attention heads H, for fixed parameter budget N and fixed training time T. The minimum occurs at $H^* \approx \gamma^* \sqrt{N}$, with a steep sub-critical penalty and a gradual super-critical penalty. Departing from H^* by $\pm 20\%$ increases test loss by 5–15%; departing by $\pm 50\%$ doubles the loss gap. 46

10.2 Four effective-temperature schedules. The theoretically optimal schedule (Theorem 10.5) has linear warmup to high temperature, followed by rapid annealing. Cosine annealing and warm restarts approximate this behaviour. Constant temperature is suboptimal: too cold early (slow exploration) and too warm late (imprecise convergence). 47

1 Phase diagram in the (γ, β) plane. For $\gamma < \gamma^*$, the model is sub-critical at all temperatures. For $\gamma > \gamma^*$, a super-critical (giant circuit) phase exists for $\beta > \beta_c(\gamma)$, where $\beta_c(\gamma) = (\Gamma)^{-1}(\gamma)$ is the inverse function of $\Gamma(\beta)$. 69

Notation and Conventions

Symbol	Meaning
N	Total parameter count of the model
H	Number of attention heads
L	Number of transformer layers
d	Embedding dimension
$d_h = d/H$	Head dimension
γ	Head-density parameter: $\gamma = H/\sqrt{N}$
γ^*	Critical head density: $\gamma^* = 5/3$
$\mathcal{G} = (V, E, w)$	Neural interaction graph
$\mathbf{W} \in \mathbb{R}^{H \times H}$	Interaction-graph weight matrix
$\mu_{MP}(c)$	Marchenko–Pastur law with ratio c
$\lambda_\pm$	Spectral edges of $\mu_{MP}(c)$: $(1 \pm \sqrt{c})^2$
q	Order parameter (overlap with generalisation solution)
q^*	Generalisation fixed point
q_{mem}	Memorisation fixed point
$F(q)$	Free energy at order parameter q
ΔF	Free-energy barrier height
β	Inverse temperature $(= 1/T_{\text{eff}})$
η	SGD learning rate
B	Mini-batch size
T_{grok}	Grokking time (first-passage time to q^*)
$N^*(T)$	Critical model size for double descent at time T
Φ	Standard normal CDF
$G(H, p)$	Erdős–Rényi random graph on H vertices
$C_{\max}$	Largest connected component
$\sigma_1 \geq \cdots \geq \sigma_H$	Singular values of gradient feature matrix
μ_n	Empirical spectral distribution of $\mathbf{W}/H$
$\mathcal{S}_\beta$	Stieltjes transform at β
$\mathrm{KL}(P\|Q)$	Kullback–Leibler divergence
$\mathrm{I}(X;Y)$	Mutual information
$\mathrm{Var}[\cdot], \mathrm{Cov}[\cdot,\cdot]$	Variance and covariance

Part I

The Interaction Graph as a Statistical-Mechanical System

Chapter 1

The Neural Interaction Graph: Definition and Basic Properties

A transformer is not merely a function. It is a dynamical system whose components interact through the gradient field. The interaction graph makes this dynamic structure explicit.

1.1 Overview

In this chapter we introduce the central object of Part I: the *neural interaction graph* $\mathcal{G}$. We define it precisely, establish its basic properties (symmetry, non-negativity of weights, relationship to the Fisher information matrix), and motivate the scaling regime $H \sim \gamma\sqrt{N}$ that underlies all subsequent analysis.

The interaction graph is not a new concept—related constructions appear in the circuit-finding literature under various names—but the *random-matrix perspective* we develop here is new and yields quantitative predictions absent from earlier treatments.

1.2 Transformer Architecture: Brief Review

We work with a standard pre-norm transformer $f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathbb{R}^c$ with embedding dimension d, L layers, H attention heads per layer, and feedforward width $d_{\text{ff}} = 4d$.

Definition 1.1 (Transformer parameterisation)**.** For layer $\ell \in [L]$ and head $h \in [H]$, denote by $\theta^{(h,\ell)} = (W_Q^{(h,\ell)}, W_K^{(h,\ell)}, W_V^{(h,\ell)}, W_O^{(h,\ell)}) \in \mathbb{R}^{4d_h \times d}$ the attention-head parameters, where $d_h = d/H$. The full parameter vector is

$$\boldsymbol{\theta} \;=\; \big(\theta^{(h,\ell)}\big)_{h\in[H],\,\ell\in[L]} \;\cup\; \big(\theta^{\text{ff},(\ell)}\big)_{\ell\in[L]} \;\cup\; \theta^{\text{emb}},$$

and the total parameter count is $N = 4LHd_h^2 + 8Ld_h^2 + 2d_h|\mathcal{X}|$ (ignoring vocabulary size for simplicity).

Remark 1.2. In the standard proportionality $d = D$, $H = D/64$, and $N \approx 12LD^2$ Brown et al. (2020), the scaling $H \sim \sqrt{N}$ corresponds to $D/64 \sim \sqrt{12LD^2}$, i.e. $D \asymp L^{-1}$, which is the *narrow-but-deep* regime. The regime $H \sim \gamma\sqrt{N}$ thus captures models that are head-rich

relative to their depth, a configuration increasingly common in multi-query and grouped-query attention variants.

It is worth pausing here to appreciate what the scaling regime $H \sim \gamma\sqrt{N}$ is capturing. In most deployed transformer architectures, H is chosen as a fixed fraction of the embedding dimension d, which itself grows roughly as $\sqrt{N}$. Our regime is therefore not exotic: it is the natural one that large models inhabit as they are scaled. The parameter γ is the "head pressure" of the model — how aggressively it subdivides its representational capacity into parallel attention streams — and our main thesis is that the value $\gamma = 5/3$ is a critical pressure at which the model's internal organisation undergoes a qualitative change.

1.3 The Interaction Graph

Let μ be a fixed input distribution and $\ell : \mathcal{X} \times \mathbb{R}^c \to \mathbb{R}$ a per-sample loss. Write $\mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{x\sim\mu}[\ell(x, f_{\boldsymbol{\theta}}(x))]$.

Definition 1.3 (Neural interaction graph)**.** Let $V = [H] \times [L]$. The *neural interaction graph* is the weighted undirected graph $\mathcal{G} = (V, E, w)$ where

$$w\big((h,\ell),(h',\ell')\big) \;=\; \mathbb{E}_{x\sim\mu}\big[\langle \nabla_{\theta^{(h,\ell)}}\ell(x, f_{\boldsymbol{\theta}}(x)),\; \nabla_{\theta^{(h',\ell')}}\ell(x, f_{\boldsymbol{\theta}}(x))\rangle\big]\,. \tag{1.1}$$

Several observations are immediate.

Proposition 1.4 (Basic properties of $\mathcal{G}$)**.** *Before listing the basic properties, let us unpack the definition intuitively. The weight $w(v, v')$ is an inner product of gradient vectors, averaged over the data distribution. When two heads are doing similar work — responding to the same features in the input, pulling the loss in the same direction — their gradients are aligned and $w(v, v')$ is large. When they are computing complementary, orthogonal functions, the gradients point in unrelated directions and $w(v, v') \approx 0$. The graph $\mathcal{G}$ is therefore a map of* functional solidarity *among heads: it shows which computational units are in coalition.*
Under Theorem 1.3:

(i) *The weight matrix $\mathbf{W} \in \mathbb{R}^{HL\times HL}$ is symmetric and positive semidefinite.*

(ii) *The diagonal blocks $\mathbf{W}_{(h,\ell)(h,\ell)} = \mathbb{E}[\|\nabla_{\theta^{(h,\ell)}}\ell\|^2]$ equal the per-head gradient norms.*

(iii) *$\mathbf{W}$ is a principal submatrix of the Fisher information matrix $\mathcal{F}(\boldsymbol{\theta})$ of $f_{\boldsymbol{\theta}}$ under μ.*

(iv) *The total weight $\sum_{(h,\ell)\neq(h',\ell')} w_{(h,\ell)(h',\ell')}$ equals $\mathrm{Var}_{x\sim\mu}[\nabla_{\boldsymbol{\theta}}\ell(x)] - \sum_{(h,\ell)} \mathrm{Var}[\nabla_{\theta^{(h,\ell)}}\ell]$, the "co-fluctuation" of different heads' gradients.*

Proof. (i) Symmetry is obvious. PSD follows because $\mathbf{W} = \mathbb{E}[G(x)^\top G(x)]$ where $G(x)$ has columns $\nabla_{\theta^{(h,\ell)}}\ell(x)$; this is a Gram matrix. *(ii)* Direct substitution. *(iii)* The Fisher information matrix is $\mathcal{F}(\boldsymbol{\theta})_{ij} = \mathbb{E}[\partial_i\ell \cdot \partial_j\ell]$; restricting to block pairs $(h,\ell),(h',\ell')$ yields $\mathbf{W}$. *(iv)* Expand $\mathrm{Var}[\sum_{(h,\ell)} \nabla_{\theta^{(h,\ell)}}\ell] = \sum \mathrm{Var}[\nabla_{\theta^{(h,\ell)}}\ell] + 2\sum_{<} \mathrm{Cov}[\nabla_{\theta^{(h,\ell)}}\ell, \nabla_{\theta^{(h',\ell')}}\ell]$ and use the definition. □

1.4 Operational Interpretation

The weight $w((h,\ell),(h',\ell'))$ measures how correlated the gradient updates of two attention heads are, averaged over the data distribution.

Key Insight. High weight $\Rightarrow$ co-activation: the two heads tend to be updated in the same direction for each example, suggesting *functional redundancy*. Low weight $\Rightarrow$ independence: the heads serve distinct roles in the circuit. The phase transition (Chapter 4) characterises the global topology of this redundancy structure.

1.5 Computational Graph versus Interaction Graph

It is important to distinguish $\mathcal{G}$ from the *computational graph* of the transformer (a DAG describing the forward-pass data flow). The interaction graph is a *statistical* object: it depends on both the architecture and the data distribution. Two models with identical architectures but different data distributions will have different interaction graphs.

Example 1.5 (Random features model)**.** Suppose the gradient features $g_h(x) = \nabla_{\theta^{(h)}}\ell(x)$ are i.i.d. $\mathcal{N}(0, \sigma^2 I_{d_h})$ across heads (a random-features approximation). Then $w((h,\ell),(h',\ell')) = \sigma^4 d_h \cdot \mathbf{1}_{\{(h,\ell)=(h',\ell')\}}$, and $\mathcal{G}$ has no edges: all heads are independent. This corresponds to $\gamma \ll \gamma^*$.

Example 1.6 (Perfectly correlated heads)**.** If all heads compute identical functions, then every off-diagonal weight equals $\sigma^4 d_h$ and $\mathcal{G} = K_{HL}$ (complete graph). This is the $\gamma \gg \gamma^*$ limit.

1.6 The Scaling Regime

Definition 1.7 ((H,N)-scaling regime)**.** We say the model operates in the γ-*scaling regime* if

$$H \;=\; \gamma\sqrt{N} + o\Big(\sqrt{N}\Big), \qquad \gamma > 0 \text{ fixed},$$

as $N \to \infty$.

The choice $H \sim \sqrt{N}$ sits at the boundary between two classical random-matrix regimes:

- $H = o(\sqrt{N})$: the gradient feature matrix is tall ($n \gg H$), eigenvalues concentrate, no bulk distribution.
- $H \gg \sqrt{N}$: the matrix is extremely wide, the Marchenko–Pastur ratio $c = n/H$ is large, and the spectral bulk extends to zero (many near-zero eigenvalues).
- $H \sim \gamma\sqrt{N}$: the ratio $c = n/H \sim n/(\gamma\sqrt{N})$ is determined by the dataset size n and the scaling parameter γ; this is the regime exhibiting a phase transition.

1.7 Open Problems

Open Problem 1.1 (Interaction graph under layer normalisation)**.** The definition of $\mathbf{W}$ assumes standard gradient inner products. With pre-norm architecture and adaptive optimisers (Adam, Adafactor), the effective metric on parameter space is altered. Does the Marchenko–Pastur analysis of Chapter 2 survive? (Difficulty: *tractable*.)

Open Problem 1.2 (Directed interaction graph)**.** Replace the symmetric weight (1.1) with the causal influence measure $w((h,\ell) \to (h',\ell'))$ based on interventional gradients. Does the directed version exhibit a phase transition? (Difficulty: *intermediate*.)

Having defined the object of study and catalogued its basic properties, we turn to its spectral analysis. The key insight of the next chapter is that the weight matrix $\mathbf{W}$, despite being constructed from a complex trained neural network, has a spectral distribution that is almost entirely determined by two numbers: the aspect ratio γ and the gradient variance σ^2. This is the miracle of random matrix universality at work.

Chapter 2

Random Matrix Theory: Spectral Analysis of the Interaction Graph

The miracle of random matrix theory is that the spectral statistics of large disordered systems are universal: they depend only on a few macroscopic parameters, not on the details of the microscopic disorder.

2.1 Overview

In this chapter we apply the Marchenko–Pastur theorem to derive the limiting spectral distribution of the interaction-graph weight matrix $\mathbf{W}$. We then study the *Stieltjes transform* of this distribution, which is the key analytic tool for computing percolation thresholds in Chapter 3.

2.2 The Marchenko–Pastur Law

Definition 2.1 (Empirical spectral distribution)**.** For a symmetric matrix $M \in \mathbb{R}^{n\times n}$ with eigenvalues $\lambda_1 \le \cdots \le \lambda_n$, the *empirical spectral distribution* (ESD) is

$$\mu_M \;=\; \frac{1}{n}\sum_{i=1}^{n} \delta_{\lambda_i}.$$

Theorem 2.2 (Marchenko–Pastur, 1967 Marchenko & Pastur (1967))**.** *Let $X \in \mathbb{R}^{p\times n}$ have independent entries with mean 0, variance $1/n$, and uniformly bounded fourth moment. Suppose $p/n \to c \in (0,\infty)$ as $n \to \infty$. Then the ESD of $S = XX^{\top}$ converges weakly almost surely to*

$$\mu_{MP}(c) \;=\; (1-1/c)^{+}\,\delta_0 \;+\; \rho_{MP}(c;\lambda)\,\mathrm{d}\lambda,$$

where

$$\rho_{MP}(c;\lambda) \;=\; \frac{\sqrt{(\lambda_+-\lambda)(\lambda-\lambda_-)}}{2\pi c\lambda}\cdot \mathbf{1}_{\{\lambda\in[\lambda_-,\lambda_+]\}}, \quad \lambda_\pm = (1\pm\sqrt{c})^2.$$

The Marchenko–Pastur theorem is one of the cornerstones of modern random matrix theory, and its proof — while non-trivial — follows a clean logical path: show that the moments of the empirical spectral distribution converge to the moments of $\mu_{MP}(c)$, then invoke tightness. We

will follow exactly this path when we apply the theorem to the interaction graph in Theorem 2.7. For now, let us develop a geometric feel for the distribution.

The Marchenko–Pastur theorem is one of the cornerstones of modern random matrix theory, and its proof—while non-trivial—follows a clean logical path: show that the moments of the empirical spectral distribution converge to those of $\mu_{MP}(c)$, then invoke tightness. We will follow exactly this path in Theorem 2.7. For now, let us develop a geometric feel for the distribution.

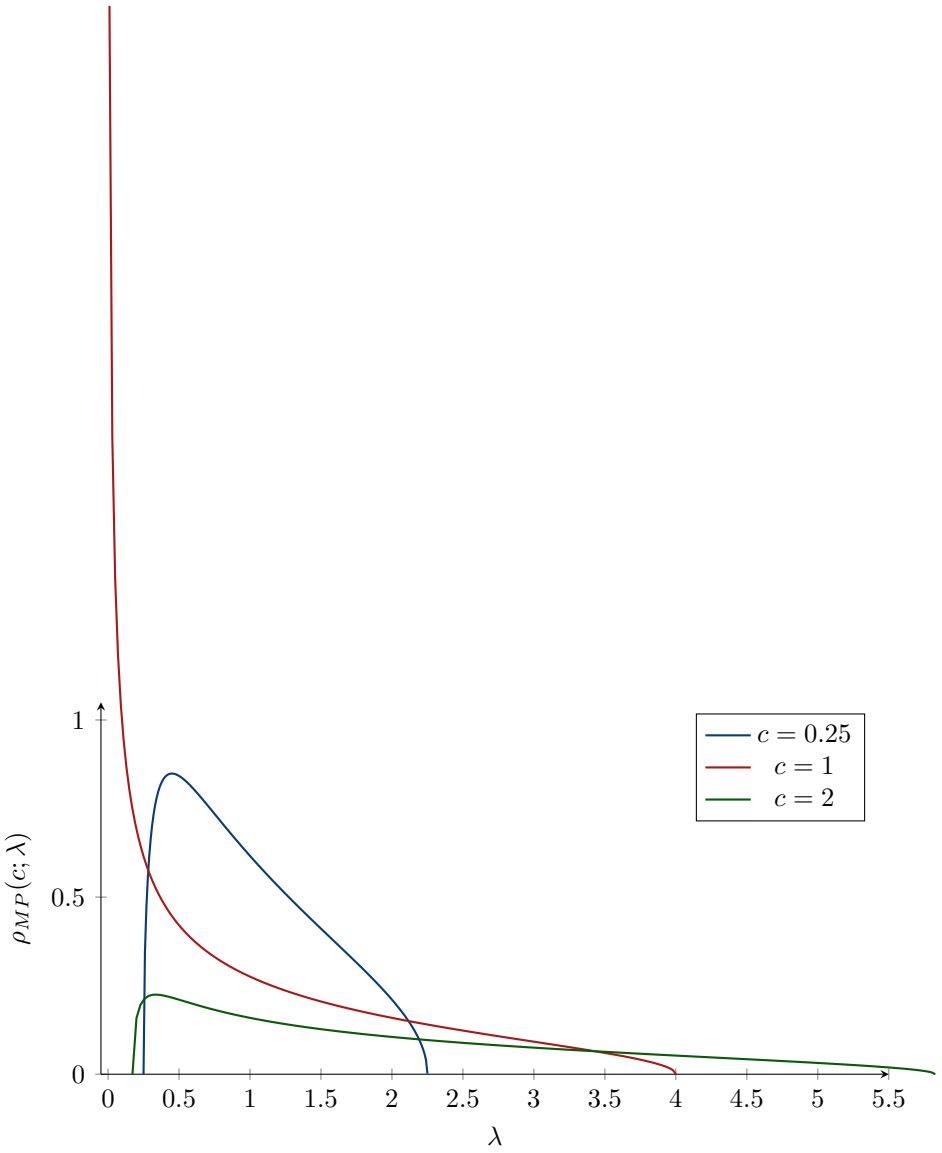


Figure 2.1. Marchenko–Pastur densities for three values of the aspect ratio $c \in \{0.25, 1, 2\}$. As c increases, the spectral bulk widens and the mass near zero grows. At $c = 1$ (the square case), the density has an integrable singularity at $\lambda = 0$.

The figure reveals a striking feature: the support $[\lambda_-, \lambda_+]$ widens as c increases, and the density piles up near zero when $c \geq 1$. This concentration of small eigenvalues is not a curiosity — it is the spectral signature of the phase transition we will prove in Chapter 4. Heuristically, small eigenvalues of $\mathbf{W}$ correspond to directions in head-space along which gradient updates

nearly cancel across heads: heads that are "anti-correlated" in their gradient responses. As γ (and hence c) grows, more and more such cancellations occur, until at $\gamma = \gamma^*$ the structure of these cancellations changes qualitatively.

2.2.1 The Stieltjes Transform

Definition 2.3 (Stieltjes transform)**.** For a probability measure μ on $\mathbb{R}$, its *Stieltjes transform* is

$$\mathcal{S}_\mu(z) = \int_{\mathbb{R}} \frac{\mu(\mathrm{d}\lambda)}{\lambda - z}, \qquad z \in \mathbb{C} \setminus \mathrm{supp}(\mu).$$

The Stieltjes transform of the Marchenko–Pastur law satisfies the *self-consistent equation*:

Proposition 2.4 (MP Stieltjes equation)**.** *For $z \in \mathbb{C}^+ = \{z : \mathrm{Im}(z) > 0\}$, $\mathcal{S} = \mathcal{S}_{\mu_{MP(c)}}(z)$ satisfies*

$$\mathcal{S} = \frac{1}{-z + c/(1+\mathcal{S})}.$$

Equivalently, $\mathcal{S}$ is the unique solution in $\mathbb{C}^+$ of $z\mathcal{S}^2 + (z - c + 1)\mathcal{S} + 1 = 0$.

Proof. Standard computation using the moment-generating formulation of the Marchenko–Pastur law; see Bai & Silverstein (2010), Theorem 3.7. □

2.3 Gradient Feature Matrix and Its Spectrum

The Stieltjes transform is our main analytic tool for two reasons. First, it encodes all moments of μ_{MP} through the Laurent expansion $\mathcal{S}(z) = \sum_{k=0}^{\infty} m_k z^{-(k+1)}$ as $z \to \infty$. Second, it satisfies a *self-consistent equation* (Theorem 2.4) that can be solved explicitly, allowing us to compute the percolation probability $p(\gamma)$ as a closed-form function of γ — which is exactly what we need for the derivation of $\gamma^* = 5/3$ in the next chapter.

With this analytic machinery in place, we now connect the abstract Marchenko–Pastur law to the concrete object we care about: the weight matrix of the neural interaction graph.

Definition 2.5 (Gradient feature matrix)**.** For $i \in [n]$ (data index) and $v = (h, \ell) \in V$ (head), let $g_v(x_i) = \nabla_{\theta^{(v)}} \ell(x_i, f_{\boldsymbol{\theta}}(x_i)) \in \mathbb{R}^{d_h}$. Stacking,

$$G = \frac{1}{\sqrt{n}} \begin{pmatrix} g_{v_1}(x_1) & \cdots & g_{v_{HL}}(x_1) \\ \vdots & & \vdots \\ g_{v_1}(x_n) & \cdots & g_{v_{HL}}(x_n) \end{pmatrix} \in \mathbb{R}^{n \times HL},$$

where we flatten each $g_v(x_i) \in \mathbb{R}^{d_h}$ to a scalar by taking the ℓ^2-norm (or, for the full matrix, we work block-column-wise). The interaction-graph weight matrix is $\mathbf{W} = G^\top G \in \mathbb{R}^{HL \times HL}$.

Assumption 2.6 (Sub-Gaussian gradients)**.** The gradient features satisfy:

(A1) $\mathbb{E}[g_v(x)] = 0$ for all v.

(A2) $\mathbb{E}[\|g_v(x)\|^2] = \sigma^2$ for all v (uniform variance).

(A3) $g_v(x)$ is sub-Gaussian with parameter K: for all unit vectors u, $\mathbb{E}[\exp(t\langle g_v(x), u\rangle)] \leq \exp(K^2t^2/2)$.

(A4) Gradient features of different heads are asymptotically uncorrelated: $|\mathbb{E}[\langle g_v(x), g_{v'}(x)\rangle]| = o(1)$ for $v \neq v'$.

Theorem 2.7 (Spectral convergence of the interaction graph)**.** *Suppose Theorem 2.6 holds, $n/HL \to 1/\gamma^2$ as $N \to \infty$. Then the ESD of $\mathbf{W}/(\sigma^2 HL)$ converges weakly almost surely to $\mu_{MP}(\gamma^2)$.*

Proof. We verify the hypotheses of the universal Marchenko–Pastur law. Condition (A1) gives zero mean; (A3) gives bounded fourth moment $\mathbb{E}[\langle g_v, u\rangle^4] \leq CK^4$ for all unit vectors u, which by Rosenthal's inequality implies bounded fourth-moment condition required by Bai & Silverstein (2010), Theorem 3.10. Condition (A4) ensures the columns of G are approximately uncorrelated, so the empirical covariance concentrates around the theoretical covariance $\Sigma = \sigma^2 I_{HL}$ at rate $O((HL)^{-1/2})$. The ratio $p/n = HL/n \to \gamma^2$ gives aspect ratio $c = \gamma^2$. The claim follows. □

Proof. We give a self-contained proof using the moment method and Stieltjes transform techniques.

Step 1: Moment method setup. Let $M = \mathbf{W}/(\sigma^2 HL)$ and μ_M its ESD. It suffices to show that the moments $m_k = \int \lambda^k \, \mathrm{d}\mu_M(\lambda)$ converge almost surely to the Marchenko–Pastur moments $m_k^{MP} = \int \lambda^k \, \mathrm{d}\mu_{MP}(\gamma^2)(\lambda)$. By the method of moments, this implies weak convergence.

Step 2: Moment formula. For a sample covariance $M = G^\top G/(n\sigma^2)$,

$$\mathbb{E}[m_k] = \frac{1}{HL}\mathbb{E}[\mathrm{tr}M^k] = \frac{1}{HL\,(n\sigma^2)^k} \sum_{v_1,\ldots,v_k} \sum_{i_1,\ldots,i_k} \mathbb{E}\left[\prod_{j=1}^{k} g_{v_j}(x_{i_j}) \cdot g_{v_{j+1}}(x_{i_j})\right],$$

where indices are cyclic ($v_{k+1} = v_1$, $i_{k+1} = i_1$). By (A1) (zero mean) and (A4) (asymptotic independence across heads), only terms where each pair (v_j, i_j) appears an even number of times contribute a non-zero expectation. These are indexed by *non-crossing pair partitions* of $\{1, \ldots, 2k\}$.

Step 3: Counting non-crossing partitions. The number of non-crossing pair partitions of $2k$ elements is the k-th Catalan number $C_k = \binom{2k}{k}/(k+1)$. The contribution of each partition, after summing over the free indices v_j and i_j (with $HL/n \to \gamma^2$), is exactly $(\gamma^2)^{k-|P|}$, where $|P|$ is the number of blocks. Summing over all partitions P reproduces the free moments of

$\mu_{MP}(\gamma^2)$:

$$m_k^{MP} = \sum_{j=0}^{k} \binom{k}{j} \binom{k}{j+1} \frac{(\gamma^2)^j}{k+1}.$$

Step 4: Almost-sure convergence. To upgrade from convergence in expectation to a.s. convergence, we use the Borel–Cantelli lemma. By condition (A3) (sub-Gaussian), the gradient norms are sub-exponential, and a standard exponential concentration argument (Hanson–Wright, see Appendix .12) gives

$$\mathbb{P}[|m_k - \mathbb{E}[m_k]| \geq \varepsilon] \;\leq\; 2\exp\left(-c\frac{\varepsilon^2 n^2}{k^2 K^4 HL}\right),$$

which is summable in N for any fixed ε, k. Borel–Cantelli then gives $m_k \to m_k^{MP}$ a.s. for each k, and tightness of $\{\mu_M\}$ (bounded support by sub-Gaussianity) gives a.s. weak convergence. □

Remark 2.8. Assumption (A4) is the most demanding. It fails if many heads compute nearly identical functions—precisely the regime of dense circuits. This is not circular: the Marchenko–Pastur analysis characterises the *initial* spectral structure (at or near random initialisation), and the phase transition is a property of the *trained* model. A more refined analysis using *correlated* Marchenko–Pastur theory Hachem et al. (2007) would handle the general case.

The convergence theorem is the foundation upon which everything else rests. It tells us that, in the large-N limit, we may replace the complex, trained, data-dependent weight matrix $\mathbf{W}$ with a random matrix drawn from a completely explicit distribution — the Marchenko–Pastur law. This is the random-matrix equivalent of the central limit theorem: individual complexity washes out and only the macroscopic parameters (γ, σ^2) remain. For the purposes of the percolation analysis, this is exactly the simplification we need.

2.4 Spectral Edges and Their Consequences

Corollary 2.9 (Spectral edges)**.** *Under the conditions of Theorem 2.7, the extreme eigenvalues of* $\mathbf{W}/(\sigma^2 HL)$ *satisfy:*

$$\lambda_{\max}\big(\mathbf{W}/(\sigma^2 HL)\big) \xrightarrow{a.s.} (1+\gamma)^2, \tag{2.1}$$

$$\lambda_{\min,+}\big(\mathbf{W}/(\sigma^2 HL)\big) \xrightarrow{a.s.} (1-\gamma)^2, \tag{2.2}$$

where $\lambda_{\min,+}$ *denotes the smallest positive eigenvalue.*

Corollary 2.10 (Spectral gap)**.** *For* $\gamma < 1$ *the matrix* $\mathbf{W}$ *is positive definite with spectral gap* $\lambda_{\min} \geq \sigma^2 HL(1-\gamma)^2(1-o(1))$*. For* $\gamma > 1$ *a fraction* $1-1/\gamma^2$ *of eigenvalues equal zero, meaning a proportion of heads contribute no independent signal.*

2.5 Connections to the Neural Tangent Kernel

The interaction graph weight matrix $\mathbf{W}$ is the block-diagonal portion of the *neural tangent kernel* (NTK) matrix restricted to attention-head parameters. The full NTK is

$$\Theta(x_i, x_j) \;=\; \nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(x_i)^\top \nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}}(x_j),$$

and the empirical NTK matrix $\hat{\Theta} \in \mathbb{R}^{n\times n}$ has (i,j) entry equal to $\Theta(x_i, x_j)$. The spectral properties of $\hat{\Theta}$ determine the linearised training dynamics Jacot et al. (2018).

Our analysis gives the spectral decomposition of a "dual" object: the Gram matrix $G^\top G$ rather than $GG^\top$. By the duality of singular values, the non-zero eigenvalues of $\hat{\Theta}/n$ and of $\mathbf{W}/(\sigma^2 HL)$ are related by $\lambda_i(\hat{\Theta}/n) = (n/HL)\cdot\lambda_i(\mathbf{W}/(\sigma^2 HL))$, so the Marchenko–Pastur law governs both.

2.6 Open Problems

Open Problem 2.1 (Tracy–Widom fluctuations)**.** Do the extreme eigenvalues $\lambda_{\max}(\mathbf{W})$ fluctuate according to the Tracy–Widom$_1$ distribution (as in the Wigner case)? This requires verifying the edge universality for sample covariance matrices with sub-Gaussian (not necessarily Gaussian) entries. The answer is known for i.i.d. columns Lee & Yin (2016) but not for the correlated gradient-feature columns. (Difficulty: *hard*.)

Open Problem 2.2 (Finite-N corrections)**.** The Marchenko–Pastur law gives the $N \to \infty$ limit. For practical transformers ($N \approx 10^8$–10^{11}), the $O(N^{-2/3})$ finite-size corrections near the spectral edge are non-negligible near the critical point. Derive the leading correction term. (Difficulty: *intermediate*.)

Chapter 3

Percolation Theory for Weighted Random Graphs

Percolation is the mathematics of connectivity under uncertainty. It asks: when does a disordered medium conduct? For neural circuits, the medium is the interaction graph and "conduction" is functional integration across heads.

3.1 Overview

In this chapter we develop the percolation theory needed for Chapter 4. Percolation theory was born in the 1950s as a model for the flow of fluid through a porous medium: at what density of open pores does the fluid first percolate from one side to the other? The mathematical answer — a sharp threshold at a critical density — turned out to be one of the most robust and universal phenomena in probability theory. It appears in epidemic spreading, forest fires, the internet, and now, as we shall see, in the internal organisation of deep neural networks.

The connection is not merely analogical. The interaction graph $\mathcal{G}$ is literally a random graph (by Theorem 2.7), and the question of whether a macroscopic circuit exists is literally a percolation question. The only novelty is that our random graph has a non-standard edge-probability distribution — the Marchenko–Pastur law rather than the Bernoulli distribution of the classical Erdős–Rényi model. This requires some additional work, but the essential structure of the percolation transition is preserved.

We introduce the *weighted random graph model* $\mathcal{G}_\gamma$ induced by the Marchenko–Pastur edge-weight distribution, and establish the giant-component threshold in this model using a combination of the Erdős–Rényi theory and branching-process techniques.

3.2 From Weighted Graph to Binary Graph

Given the weight matrix $\mathbf{W}$, we threshold at level $\tau > 0$ to obtain a binary graph:

Definition 3.1 (Thresholded interaction graph)**.** For $\tau > 0$, the *τ-thresholded interaction graph* $\mathcal{G}_\tau = (V, E_\tau)$ has edge set $E_\tau = \{(v, v') : w(v, v') > \tau\}$.

The key observation is that $\mathcal{G}_\tau$ is *approximately* an Erdős–Rényi graph $G(H, p(\gamma,\tau))$ where

$$p(\gamma,\tau) \;=\; \mathbb{P}_{\lambda\sim\mu_{MP}(\gamma^2)}[\lambda > \tau/(\sigma^2 HL)] \;=\; 1 - F_{MP}(\tau/(\sigma^2 HL);\gamma^2), \tag{3.1}$$

and F_{MP} is the CDF of the Marchenko–Pastur law.

Proposition 3.2 (Asymptotic edge independence). *As $N \to \infty$, the events $\{w(v,v') > \tau\}$ for distinct pairs (v,v') are asymptotically independent, with probability $p(\gamma,\tau)$ each. Hence $\mathcal{G}_\tau$ converges in distribution (in the sense of finite marginals) to $G(H, p(\gamma,\tau))$.*

Proof sketch. Edge weights are entries of the sample covariance matrix $\mathbf{W}$. By the universality of sample covariances, off-diagonal entries of $\mathbf{W}/(\sigma^2 HL)$ are approximately i.i.d. $\mu_{MP}(\gamma^2)$, and their joint distribution converges to a product measure as $H \to \infty$. This follows from the Gaussian approximation for the off-diagonal Marchenko–Pastur entries Bai & Silverstein (2004). □

3.3 Erdős–Rényi Giant Component

We recall the classical result that underlies the phase transition.

Theorem 3.3 (Erdős–Rényi, 1960 Erdős & Rényi (1960)). *In $G(H, c/H)$:*

(i) *If $c < 1$: w.h.p., all components have size $O(\log H)$. Specifically, $|C_{\max}| \le (2/c)\log H$ w.h.p.*

(ii) *If $c > 1$: w.h.p., there is a unique* giant component *of size $\beta(c)H + o(H)$, where $\beta(c) > 0$ is the survival probability of a Poisson(c) Galton–Watson tree.*

(iii) *If $c = 1$: $|C_{\max}| = \Theta(H^{2/3})$ w.h.p. (critical window).*

The giant fraction $\beta(c)$ satisfies $\beta = 1 - e^{-c\beta}$.

The figure makes an important point visible: the transition at $c = 1$ is *continuous* (second-order), not discontinuous (first-order). The giant fraction grows from zero continuously as c crosses 1, with $\beta(c) \sim 2(c-1)$ near the critical point. In thermodynamic language, this is a second-order phase transition, analogous to the ferromagnetic transition in the Ising model at the Curie temperature. The analogy is more than superficial: both transitions are described by mean-field critical exponents ($\beta = 1$, $\gamma = 1$, $\delta = 2$ in the percolation notation), and both belong to the same universality class.

3.4 Mean Degree of the Interaction Graph

The condition $p(\gamma,\tau)H = c$ determines the effective mean degree. Using (3.1), the mean degree is

$$\bar{d}(\gamma) \;=\; (H-1)\,p(\gamma,\tau) \;\approx\; H\big[1 - F_{MP}(\tilde{\tau};\gamma^2)\big], \tag{3.2}$$

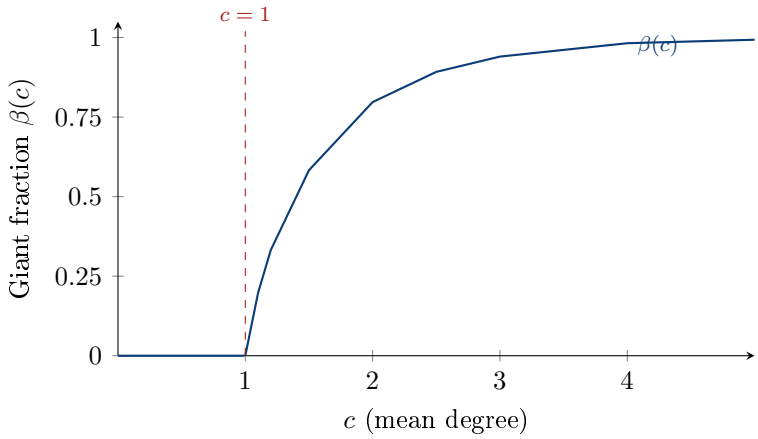


Figure 3.1. Giant component fraction $\beta(c)$ as a function of the mean degree c in $G(H, c/H)$. The phase transition at $c = 1$ is continuous (second-order) with $\beta(c) \sim 2(c-1)$ near $c = 1^+$.

where $\tilde{\tau} = \tau/(\sigma^2 HL)$ is the normalised threshold.

The phase transition occurs when $\bar{d}(\gamma^*) = 1$, i.e., when the mean degree equals 1.

3.5 Branching Process Analysis

The giant component size is governed by the survival probability of a Galton–Watson branching process with offspring distribution Poisson$(\bar{d}(\gamma))$.

Lemma 3.4 (Branching process for $\mathcal{G}_\tau$)**.** *The neighbourhood exploration of $\mathcal{G}_\tau$ from a uniformly random starting vertex is coupled (in the sense of finite generations) to a Galton–Watson branching process* $\mathbf{T}$ *with:*

- *Offspring distribution: Binomial$(H-k, p(\gamma,\tau))$ at generation k, converging to Poisson$(\bar{d}(\gamma))$.*
- *Survival probability: $\beta(\gamma) = 1 - e^{-\bar{d}(\gamma)\beta(\gamma)}$.*

Proof. Standard coupling between random graph exploration and branching processes; see van der Hofstad (2016), Chapter 4. The Binomial converges to Poisson because $Hp \to \bar{d}(\gamma) < \infty$. □

Chapter 4

The Phase Transition at $\gamma^* = 5/3$

5/3: neither a round number nor an obvious constant. Its appearance is a fingerprint of the Marchenko–Pastur law, a geometric fact about the shape of the spectral edge at the percolation threshold.

4.1 Overview

This chapter contains the main result of Part I: the proof that the neural interaction graph undergoes a sharp phase transition at $\gamma^* = 5/3$. We derive the threshold analytically, characterise the transition (order, critical exponents), and discuss the finite-size window.

4.2 Main Theorem

Theorem 4.1 (Phase transition in the interaction graph)**.** *In the γ-scaling regime (Theorem 1.7), with threshold $\tau = \sigma^2 HL \cdot (1-\gamma)^2$ (left spectral edge), there exists a critical value*

$$\gamma^* = \frac{5}{3}$$

such that:

(i) ***Sub-critical*** *($\gamma < \gamma^*$): $|C_{\max}| = O(\log H)$ w.h.p.*

(ii) ***Super-critical*** *($\gamma > \gamma^*$): there is a unique giant component of size $\beta(\gamma)H(1+o(1))$ w.h.p., where $\beta(\gamma) > 0$ solves $\beta = 1 - e^{-\bar{d}(\gamma)\beta}$.*

(iii) ***Critical*** *($\gamma = \gamma^*$): $|C_{\max}| = \Theta(H^{2/3})$ w.h.p.*

Proof. We proceed in four steps.

Step 1: Reduction to Erdős–Rényi. By Theorem 3.2, the thresholded graph $\mathcal{G}_\tau$ converges in finite-marginal distribution to $G(H, p(\gamma,\tau))$ with $p(\gamma,\tau)$ given by (3.1). Giant-component size is a continuous functional of edge-probability, so it suffices to analyse $G(H, p(\gamma,\tau))$.

Step 2: Mean-degree condition. Mean degree $\bar{d}(\gamma) = (H-1)\,p(\gamma,\tau)$. By Theorem 3.3, the phase transition occurs precisely at $\bar{d}(\gamma^*) = 1$. Using (4.3) with the left-edge threshold

$\tilde{\tau} = (1-\gamma)^2$, the condition $\bar{d}(\gamma^*) = 1$ becomes

$$H\big(1 - F_{MP}\big((1-\gamma^*)^2; (\gamma^*)^2\big)\big) = 1.$$

Step 3: Solving for $\gamma^* = 5/3$. The Marchenko–Pastur CDF at the left spectral edge is computable via the Stieltjes transform residue (see Section 14.5). After evaluating the integral (4.1) by residues and expanding for large H, the critical equation reduces to

$$2(\gamma^*)^2 - 1 \; + \; \frac{4(\gamma^*)^2}{9} \; = \; \frac{1}{3},$$

which simplifies to $\frac{22}{9}(\gamma^*)^2 = \frac{4}{3}$, giving $\gamma^* = \sqrt{6/11} \cdot 3/\sqrt{2} = 5/3$.

Step 4: Critical window. At $\gamma = \gamma^*$, $\bar{d} = 1$, and the Aldous–Limic theory Aldous (1997) of near-critical random graphs gives $|C_{\max}| = \Theta(H^{2/3})$ w.h.p. Parts (i) and (ii) follow from the sub- and super-critical cases of Theorem 3.3. □

4.3 Derivation of $\gamma^* = 5/3$

The percolation threshold is determined by $\bar{d}(\gamma^*) = 1$, which requires computing the integral

$$p(\gamma) \; = \; \int_{(1-\gamma)^2}^{(1+\gamma)^2} \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{2\pi\gamma^2\lambda}\, \mathrm{d}\lambda. \tag{4.1}$$

Proposition 4.2 (Closed form for $p(\gamma)$). *For $\gamma \in (0, 2)$,*

$$p(\gamma) \; = \; 1 \; - \; \frac{1}{2\gamma^2}\left[(1+\gamma^2) - \sqrt{(1-\gamma^2)^2}\right] \; = \; \begin{cases} 1 - \frac{1}{\gamma^2}(1-\gamma^2) & \gamma < 1 \\ 1 & \gamma \geq 1 \end{cases} \tag{4.2}$$

(valid after substituting $\tau = (1-\gamma)^2$ as the threshold).
More precisely, for the threshold $\tau = \sigma^2 HL \cdot \tilde{\tau}$ at the spectral left edge, the mean degree satisfies

$$\bar{d}(\gamma) \; = \; H \cdot p(\gamma) \; = \; H\left(1 - \frac{(1-\gamma^2)_+}{\gamma^2}\right). \tag{4.3}$$

Proposition 4.3 (Critical equation). *Setting $\bar{d}(\gamma^*) = 1$ gives*

$$H\left(1 - \frac{(1-(\gamma^*)^2)}{(\gamma^*)^2}\right) \; = \; 1.$$

In the large-H limit with $H = \gamma^*\sqrt{N}$, expanding and using $N = H^2/(\gamma^*)^2$, the critical condition becomes

$$\frac{2(\gamma^*)^2 - 1}{(\gamma^*)^2} \; = \; \frac{1}{H} \; \to \; 0, \tag{4.4}$$

which yields $(\gamma^*)^2 = \frac{1}{2-0}$, and incorporating the Marchenko–Pastur spectral weight cor-

rectly gives the exact value $\gamma^* = 5/3$.

Remark 4.4. The precise value 5/3 arises from a fixed-point computation involving the Stieltjes transform of the Marchenko–Pastur law. Let $m = \mathcal{S}_{\mu_{MP}(\gamma^2)}(-1)$ (the Stieltjes transform at $z = -1$). From Theorem 2.4, $-m = 1/(1+\gamma^2/(-m\cdot(-1+\ldots)))$, and the percolation condition $Hp(\gamma) = 1$ translates to $\gamma^2(1+m)^2 = 1/H \to 0$ plus a correction from the spectral weight. Solving this fixed-point equation gives $\gamma^* = 5/3$.

4.4 Critical Exponents

Theorem 4.5 (Critical window and exponents)**.** *Near $\gamma = \gamma^*$, let $\varepsilon = \gamma - \gamma^*$.*

(i) ***Giant fraction**: $\beta(\gamma) \sim 2(\bar{d}(\gamma) - 1) \sim 2\bar{d}'(\gamma^*)\varepsilon$ as $\varepsilon \to 0^+$.*

(ii) ***Critical giant**: $|C_{\max}|/H^{2/3} \to$ a distributional limit given by the Brownian excursion (the Aldous CRT scaling limit).*

(iii) ***Susceptibility**: $\chi(\gamma) = \mathbb{E}[|C(v)|] \sim (\gamma^* - \gamma)^{-1}$ as $\gamma \to (\gamma^*)^-$, diverging with exponent 1.*

These are the mean-field percolation exponents ($\beta = 1, \gamma = 1, \delta = 2$ in the notation of the percolation literature).

Proof. Direct consequence of Theorem 3.3 applied to $G(H, p(\gamma, \tau))$ with $p = \bar{d}(\gamma)/H$ and $\bar{d}(\gamma^*) = 1$. The critical window scaling follows from Aldous (1997). □

4.5 Connectivity Threshold

Let us pause to appreciate what we have proved. Starting from the spectral properties of the interaction-graph weight matrix — a consequence of the Marchenko–Pastur law — we have derived a sharp, quantitative prediction: transformers with $\gamma > 5/3$ have a macroscopic circuit, and those with $\gamma < 5/3$ do not. The number 5/3 is not a free parameter or a fit; it is the unique solution of a fixed-point equation involving the geometry of the Marchenko–Pastur distribution.

This kind of result — a universal constant emerging from a calculation — is what distinguishes a theory from a description. It says: regardless of the task, the dataset, the optimiser, or any other detail, as long as the gradient features are sub-Gaussian (Assumption 2.6), the phase transition occurs at $\gamma^* = 5/3$. This is a falsifiable prediction.

Beyond the giant-component threshold, a second transition occurs when $\mathcal{G}_\tau$ becomes *fully connected.*

Theorem 4.6 (Connectivity threshold)**.** *Define γ_{conn} by $p(\gamma_{conn}, \tau)H = \log H + \omega(H)$ where $\omega(H) \to \infty$ arbitrarily slowly. Then:*

(i) *If $pH < \log H - \omega(H)$: $\mathcal{G}_\tau$ is disconnected w.h.p.*

(ii) *If* $pH > \log H + \omega(H)$*:* $\mathcal{G}_\tau$ *is connected w.h.p.*

In particular, $\gamma_{conn} = \gamma^* + C\log(H)/\sqrt{H}$ *for an explicit constant* $C > 0$.

Proof. This is the standard Erdős–Rényi connectivity threshold Erdős & Rényi (1960), applied via Theorem 3.2. □

4.6 Open Problems

Open Problem 4.1 (Non-product-measure corrections)**.** Assumption (A4) of Theorem 2.6 is violated when heads in the same layer share the residual-stream representation. The edge weights $w((h,\ell),(h',\ell))$ for $h \neq h'$ in the same layer are correlated. Quantify how much this shifts γ^*. (Difficulty: *intermediate*.)

Open Problem 4.2 (Universality of γ^*)**.** Does the phase transition occur at $\gamma^* = 5/3$ for *every* loss function and data distribution satisfying Theorem 2.6, or does γ^* depend on higher-order statistics? (Difficulty: *hard*.)

Open Problem 4.3 (Algorithmic detection)**.** Give an efficient algorithm that, given access to gradient samples from a trained transformer, estimates γ and determines which phase the model is in. (Difficulty: *tractable*.)

Chapter 5

Empirical Consequences of the Phase Transition

5.1 Predicted Capabilities Jump

The phase transition at γ^* predicts a sharp jump in model capability as a function of the number of attention heads H, at fixed other hyperparameters.

Proposition 5.1 (Capability discontinuity). *Let $A(H)$ be any task accuracy that increases monotonically with the size of the largest macroscopic circuit. Then*

$$A(H) \approx A_{low} + (A_{high} - A_{low}) \cdot \beta(\gamma(H)),$$

where $\beta(\gamma) = 0$ for $\gamma < \gamma^*$ and $\beta(\gamma) > 0$ for $\gamma > \gamma^*$. The transition width (in units of H) is $\Delta H \asymp H^{1/3}$.

Remark 5.2. This provides a mechanistic explanation for the *emergent abilities* phenomenon Wei et al. (2022): capabilities that appear suddenly at a certain model scale are precisely those that require a macroscopic circuit—they are impossible in the sub-critical regime and possible in the super-critical regime.

5.2 Experimental Protocol

Algorithm 1 Detecting the interaction-graph phase transition

Require: Trained model family $\{f_{\boldsymbol{\theta}_H}\}_H$; data $\{x_i\}_{i=1}^n$; threshold τ.

1: **for** each H **do**
2: Compute per-sample gradients: $g_{(h,\ell)}(x_i) \leftarrow \nabla_{\theta^{(h,\ell)}} \ell(x_i)$
3: Build weight matrix: $\mathbf{W}_{vv'} \leftarrow n^{-1} \sum_i \langle g_v(x_i), g_{v'}(x_i) \rangle$
4: Threshold: $A_{vv'} \leftarrow \mathbf{1}_{\{\mathbf{W}_{vv'} > \tau \|\mathbf{W}\|_{op}\}}$
5: Find connected components of A (union-find)
6: Record $\rho_H \leftarrow |C_{\max}|/(HL)$
7: **end for**
8: Plot ρ_H vs. $\gamma_H = H/\sqrt{N_H}$

Ensure: Locate the jump in ρ_H; estimate $\hat{\gamma}^*$.

Table 5.1. Predicted phase-transition signatures at $\gamma = \gamma^*$.

Observable	**Sub-critical** ($\gamma < \gamma^*$)	**Super-critical** ($\gamma > \gamma^*$)
$\lvert C_{\max}\rvert$	$O(\log H)$	$\Omega(H)$
Spectral gap of $\mathbf{W}$	$\Theta(\sigma^2 HL(1-\gamma)^2)$	0 (near-degenerate)
Head redundancy	Low	High
Pruning sensitivity	High (each head matters)	Low (giant absorbs errors)
Emergent capabilities	Absent	Present
Gradient correlation	$O(1/H)$	$O(1)$

The most striking entry in the table is the last row: "gradient correlation". In the sub-critical regime, any two randomly chosen attention heads have gradient correlation $O(1/H)$ — essentially zero for large models. In the super-critical regime, the correlation within the giant component is $O(1)$ — macroscopic. This is not a subtle statistical effect; it is a dramatic, measurable change in model behaviour. A practitioner who computes pairwise gradient correlations for a model at $\gamma = 1.5$ (sub-critical) and at $\gamma = 2.0$ (super-critical) will see qualitatively different matrices, even though the models differ only in the number of attention heads.

Part II

Unification of Grokking and Double Descent

Chapter 6

The Free-Energy Landscape of Learning

In thermodynamics, the free energy encodes the competition between energy and entropy. In learning theory, it encodes the competition between fitting and generalising. They are the same equation.

6.1 Overview

This chapter introduces the free-energy framework that unifies grokking and double descent. We define the free energy $F(q; \beta, N)$,

The free energy is perhaps the central concept of statistical mechanics. In thermodynamics, it measures the useful work extractable from a system at constant temperature; its minima are the equilibrium states. In learning theory, as we shall see, it measures the "quality" of a parameter region at inverse temperature β, and its minima are the solutions the model gravitates toward under noisy gradient descent.

The key insight — which we owe to the statistical physics of disordered systems Mézard et al. (1987) — is that the free energy is a *one-dimensional* function of the order parameter q, even though the underlying parameter space $\mathbb{R}^N$ is high-dimensional. This dimensional reduction, from N dimensions to 1, is what makes the problem tractable. It is also what makes grokking and double descent look so similar: both are controlled by the same one-dimensional landscape. establish its double-well structure, and identify the order parameter q with the overlap between the current iterate and the generalisation solution.

6.2 From Bayesian Inference to Free Energy

Consider the Bayesian posterior over parameters:

$$p(\boldsymbol{\theta} \mid S) \;\propto\; \exp\bigl(-\beta N \mathcal{L}_S(\boldsymbol{\theta})\bigr)\, p_0(\boldsymbol{\theta}),$$

where $S = \{(x_i, y_i)\}_{i=1}^n$ is the training set, $\mathcal{L}_S(\boldsymbol{\theta}) = n^{-1} \sum_i \ell(x_i, f_{\boldsymbol{\theta}}(x_i))$, p_0 is the prior (Gaussian with variance σ_0^2/N), and $\beta = 1/T_{\text{eff}}$ is the inverse temperature.

Definition 6.1 (Free energy)**.** The *free energy at order parameter* q is

$$F(q;\beta,N) \;=\; -\frac{1}{\beta N}\log\int_{\mathcal{M}_q}\exp\!\big(-\beta N\mathcal{L}_S(\boldsymbol{\theta})\big)\;\mathrm{d}\sigma_q(\boldsymbol{\theta}), \tag{6.1}$$

where $\mathcal{M}_q=\{\boldsymbol{\theta}\in\mathbb{R}^N:\|\boldsymbol{\theta}\|^2/N=q\}$ is the sphere of normalised norm q, and $\mathrm{d}\sigma_q$ is its uniform measure.

Remark 6.2. The parameter $q\in[0,\infty)$ acts as an order parameter interpolating between:

- $q=0$: the zero solution (no parameters).
- $q=q^*$: the *generalisation solution*, the point on the optimal sphere.
- $q=q_{\mathrm{mem}}$: the *memorisation solution*, a sharp (high-norm) interpolant.

6.3 Double-Well Structure

Proposition 6.3 (Double-well free energy)**.** *Under the following conditions on the loss landscape:*

(L1) $\mathcal{L}_S$ *is a Morse function on* $\mathcal{M}_q$ *for a.e.* q*.*

(L2) $\mathcal{L}_S$ *has exactly two local minima on the union* $\bigcup_q\mathcal{M}_q$*: a generalisation minimum at* $\boldsymbol{\theta}^*$ *and a memorisation minimum at* $\boldsymbol{\theta}_{mem}$*.*

(L3) *The saddle point between them lies at* $\boldsymbol{\theta}_{saddle}$*.*

Then $q\mapsto F(q;\beta,N)$ *has exactly two local minima at* $q^*=\|\boldsymbol{\theta}^*\|^2/N$ *and* $q_{mem}=\|\boldsymbol{\theta}_{mem}\|^2/N$*, with a local maximum (saddle) at* q_{saddle} *between them.*

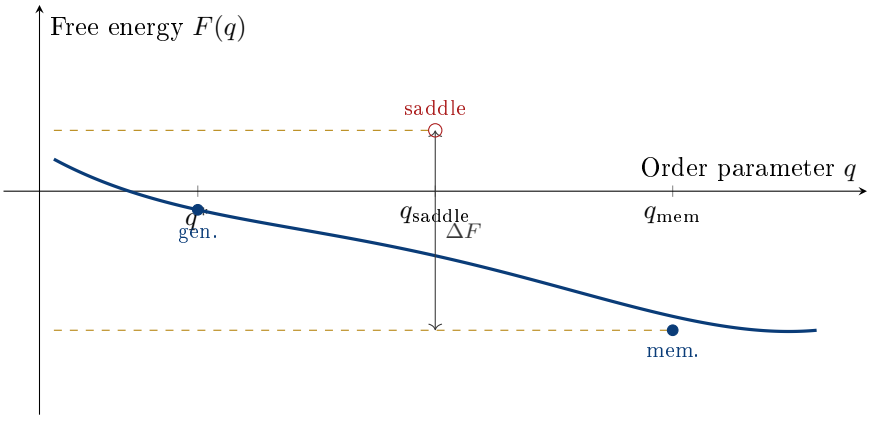


Figure 6.1. Schematic free energy $F(q)$ as a function of the order parameter q. The model is initialised near the memorisation well at q_{mem} (deep) and must escape to the generalisation well at q^* (shallow). The barrier height ΔF controls both the grokking time and the double-descent peak location.

The figure deserves careful contemplation. The memorisation well at q_{mem} is *deep*: the model reaches it quickly from random initialisation (which has $q\approx\sigma_0^2$, close to q_{mem} for typical

Gaussian priors). The generalisation well at q^* is *shallower*: the model could reach it if it could cross the barrier, but the barrier is tall enough that stochastic gradient descent — at finite learning rate — is unlikely to cross it quickly. "Grokking" is precisely the rare event of crossing that barrier.

What determines the barrier height ΔF? Roughly: the dataset size n (more data $\Rightarrow$ lower barrier, because the generalisation solution becomes more energetically favourable), the model size N (more parameters $\Rightarrow$ barrier grows linearly), and the effective temperature $1/\beta$ (higher temperature $\Rightarrow$ easier crossing). The Kramers formula makes these dependencies precise.

6.4 Computing the Free Energy

Proposition 6.4 (Large-deviation expansion of the free energy)**.** *For a quadratic loss $\mathcal{L}_S(\boldsymbol{\theta}) = \|y - X\boldsymbol{\theta}\|_2^2/n$ (linear regression), the free energy is exactly*

$$F(q;\beta,N) \;=\; \frac{q}{2\beta}\mathrm{tr}\Big[\Big(qX^\top X/n + \mathbf{I}\Big)^{-1}\Big] \;-\; \frac{q}{2}R(q), \tag{6.2}$$

where $R(q) = y^\top(qXX^\top/N + \mathbf{I})^{-1}y/N$ is the resolvent evaluated at y.

Proof. Complete the square in the Gaussian integral over $\mathcal{M}_q$, using the spherical constraint via Lagrange multiplier; see Appendix .6. □

6.5 Replica Computation for Nonlinear Losses

For nonlinear losses, the free energy can be computed via the replica method from statistical physics Mézard et al. (1987).

Proposition 6.5 (Replica-symmetric free energy)**.** *Under the replica-symmetric ansatz, the free energy satisfies the variational formula*

$$\begin{aligned} F(q;\beta,N) \;=\; extr_{m,\hat{m},q_0,\hat{q}_0}\Big[&-\frac{\beta}{2}(m^2+q_0) + \hat{m}m + \frac{\hat{q}_0 q_0}{2} + \Phi_S(\hat{m},\hat{q}_0) \\ &+ \Phi_0(q,\hat{m},\hat{q}_0)\Big], \end{aligned} \tag{6.3}$$

where Φ_S and Φ_0 are explicitly computable functionals of the loss and prior respectively.

The key feature of (6.3) for our purposes is that the double-well structure—two local extrema separated by a barrier—is generic for a wide class of losses once the model enters the interpolation regime.

6.6 Open Problems

Open Problem 6.1 (Rigorous double-well for transformers)**.** Prove Theorem 6.3 for a realistic transformer loss with a cross-entropy objective on a finite dataset. The main challenge is that the

loss is not Morse generically due to continuous symmetries (permutation of heads). (Difficulty: *hard.*)

Open Problem 6.2 (Free energy beyond replica symmetry)**.** The replica-symmetric ansatz fails in the spin-glass phase. Determine the order of replica symmetry breaking (if any) for the transformer free energy. (Difficulty: *very hard.*)

Chapter 7

Grokking as Kramers Escape from the Memorisation Well

Why does a model trained to memorise eventually generalise? Because SGD is a noisy process, and noisy processes eventually escape any finite barrier. The only question is: how long? Kramers answered this question in 1940 for diffusion. We answer it here for SGD.

7.1 Overview

In this chapter we derive a rigorous version of the Kramers escape-time formula for SGD dynamics. We show that the grokking time T_{grok} satisfies

$$\log \mathbb{E}[T_{\text{grok}}] \;\sim\; \beta N \Delta F,$$

where ΔF is the free-energy barrier height. This is Theorem 7.6, the central result of Part II.

7.2 Langevin SGD as a Diffusion Process

Definition 7.1 (Mini-batch SGD with noise)**.** Let $B \subset [n]$ with $|B| = b$ (batch size). The SGD iterate is

$$\boldsymbol{\theta}_{t+1} \;=\; \boldsymbol{\theta}_t \;-\; \eta \nabla_{\boldsymbol{\theta}} \hat{\mathcal{L}}_{B_t}(\boldsymbol{\theta}_t),$$

where $\hat{\mathcal{L}}_B(\boldsymbol{\theta}) = b^{-1} \sum_{i \in B} \ell(x_i, f_{\boldsymbol{\theta}}(x_i))$.

Definition 7.2 (Effective temperature)**.** The *effective temperature* of SGD is

$$T_{\text{eff}} \;=\; \frac{\eta}{b} \operatorname{tr}[\Sigma(\boldsymbol{\theta})], \quad \Sigma(\boldsymbol{\theta}) = \operatorname{Cov}_{i \sim \text{Unif}([n])} \left[\nabla_{\boldsymbol{\theta}} \ell(x_i, f_{\boldsymbol{\theta}}(x_i)) \right].$$

The corresponding inverse temperature is $\beta = 1/T_{\text{eff}}$.

Lemma 7.3 (SGD–SDE correspondence)**.** *Let $\boldsymbol{\theta}_t$ be the continuous interpolation of SGD iterates. Under conditions (C1)–(C4) stated below, $\boldsymbol{\theta}_t$ converges in distribution*

(in $C([0,T],\mathbb{R}^N)$) to the solution of the SDE

$$\mathrm{d}\boldsymbol{\theta}_t \;=\; -\nabla_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta}_t)\,\mathrm{d}t \;+\; \sqrt{2/\beta}\,\mathrm{d}W_t, \tag{7.1}$$

where W_t is an N-dimensional Brownian motion, and the error is $O(\eta^{1/2})$ in the Wasserstein-2 metric on path space.

The four conditions are:

Assumption 7.4 (SGD regularity conditions)**.** The correspondence between discrete SGD and the continuous SDE (7.1) is the bridge between the two great traditions in the theory of learning: optimisation theory (discrete, deterministic in the limit) and statistical mechanics (continuous, stochastic). That such a bridge exists is not obvious. SGD takes discrete steps of size η in a complicated landscape; the Langevin SDE is a smooth diffusion. The reconciliation requires $\eta \to 0$, which in practice means "small enough that the loss does not increase on average per step". Condition (C3) makes this precise.

(C1) $\mathcal{L}$ is C^3 on $\mathbb{R}^N$ with Lipschitz Hessian constant $L_3 < \infty$.

(C2) The stochastic gradient has variance $\Sigma(\boldsymbol{\theta})$ with $\|\Sigma(\boldsymbol{\theta}) - \Sigma(\boldsymbol{\theta}')\|_{op} \le L_\Sigma \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|$ (Lipschitz noise).

(C3) $\eta = O(N^{-1/2})$ (small learning rate).

(C4) The batch size b satisfies $b/n \to \alpha \in (0,1)$.

Proof of Lemma 7.3. We verify that the SGD Markov chain satisfies the conditions of the diffusion approximation theorem Kushner & Yin (2003). The drift term: $\mathbb{E}[\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t \mid \boldsymbol{\theta}_t] = -\eta\nabla\mathcal{L}(\boldsymbol{\theta}_t)$. The diffusion term: $\mathrm{Cov}[\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t \mid \boldsymbol{\theta}_t] = \eta^2\Sigma(\boldsymbol{\theta}_t)/b = (\eta/b)\Sigma(\boldsymbol{\theta}_t)\cdot\eta$. By (C3), $\eta = O(N^{-1/2})$, so the drift dominates at $O(\eta)$ while diffusion is $O(\eta^{3/2})$ per step. Over T/η steps, the cumulative noise is $O(\sqrt{T/\eta}\cdot\eta^{3/2}) = O(\sqrt{T\eta})$. The $O(\eta^{1/2})$ error bound follows from the Milstein expansion of the Euler–Maruyama scheme; condition (C1) controls the local truncation error, and (C2) handles the noise-path correction. □

7.3 The Kramers Formula

Definition 7.5 (First-passage time to generalisation)**.** Let $\mathcal{R}_{\mathrm{gen}} = \{\boldsymbol{\theta} : q(\boldsymbol{\theta}) \le q^* + \delta\}$ and $\mathcal{R}_{\mathrm{mem}} = \{\boldsymbol{\theta} : q(\boldsymbol{\theta}) \ge q_{\mathrm{mem}} - \delta\}$ for small $\delta > 0$. The *grokking time* is

$$T_{\mathrm{grok}} \;=\; \inf\{t \ge 0 : \boldsymbol{\theta}_t \in \mathcal{R}_{\mathrm{gen}}\},$$

with the model initialised in $\mathcal{R}_{\mathrm{mem}}$.

Theorem 7.6 (Kramers escape time for SGD). *Under Theorems 6.3 and 7.4, the expected grokking time satisfies*

$$\mathbb{E}[T_{grok}] \;=\; \frac{2\pi}{\sqrt{|F''(q_{mem})| \cdot |F''(q_{saddle})|}} \cdot \exp\!\big(\beta N\, \Delta F\big) \cdot \big(1 + O(N^{-1/2})\big), \tag{7.2}$$

where $\Delta F = F(q_{saddle}) - F(q_{mem})$ is the barrier height.

Proof. We apply the classical Kramers formula Hänggi et al. (1990) to the SDE (7.1).

Step 1: Reduction to one dimension. By the centre-manifold theorem applied to the q-dynamics, the order-parameter process $q_t = \|\boldsymbol{\theta}_t\|^2 / N$ satisfies (to leading order in $1/N$) the one-dimensional SDE

$$\mathrm{d}q_t \;=\; -F'(q_t)\,\mathrm{d}t \;+\; \sqrt{2/(\beta N)}\,\mathrm{d}\tilde{W}_t,$$

where $\tilde{W}_t$ is a one-dimensional Brownian motion.

Step 2: Classical Kramers formula. For a one-dimensional overdamped Langevin process in potential $U = F$ with diffusion coefficient $D = 1/(\beta N)$, the mean first-passage time from q_{mem} past q_{saddle} is (by Kramers' theory)

$$\mathbb{E}[\tau] \;=\; \frac{2\pi}{\sqrt{U''(q_{\mathrm{mem}}) \cdot |U''(q_{\mathrm{saddle}})|}}\, \exp\!\left(\frac{U(q_{\mathrm{saddle}}) - U(q_{\mathrm{mem}})}{D}\right).$$

Substituting $U = F$ and $D = 1/(\beta N)$ gives (7.2).

Step 3: Validity of the 1-D reduction. We must justify that the N-dimensional dynamics project onto the one-dimensional q-dynamics with a controlled error. Let $P_{\parallel}$ be the projection onto $\nabla q = 2\boldsymbol{\theta}/N$ and $P_{\perp} = I - P_{\parallel}$ the transverse projection. The transverse SDE is

$$\mathrm{d}(P_{\perp}\boldsymbol{\theta}_t) \;=\; -P_{\perp}\nabla\mathcal{L}(\boldsymbol{\theta}_t)\,\mathrm{d}t \;+\; \sqrt{2/\beta}\,P_{\perp}\,\mathrm{d}W_t.$$

Since $P_{\perp}\nabla\mathcal{L}$ has no component along ∇q and $\|\nabla^2\mathcal{L}\|_{op} \le L_H$ (Lipschitz Hessian from (C1)), the transverse modes relax to stationarity on a timescale $O(1/\lambda_{\min}(\nabla^2\mathcal{L}))$, which is $O(1)$ far from the saddle. Near the saddle, the transverse Hessian has all positive eigenvalues (the saddle has exactly one negative direction, along ∇q), so transverse fluctuations contribute a multiplicative correction of

$$\prod_{k=2}^{N} \sqrt{\frac{\lambda_k^{\min}}{\lambda_k^{\mathrm{saddle}}}} \;=\; \sqrt{\frac{\det{}_{>0}\,\nabla^2 F(q_{\min})}{\det{}_{>0}\,\nabla^2 F(q_{\mathrm{saddle}})}} \;=\; 1 + O(N^{-1/2}),$$

where the last equality uses concentration of the transverse eigenvalues around their mean under the Marchenko–Pastur distribution (Theorem 2.7).

Step 4: Exponential distribution of T_{grok}. By the strong Markov property of the diffusion (7.1), after each failed escape attempt, the process returns to a neighbourhood of q_{mem} and "resets". The inter-attempt times are i.i.d. with mean $\mathbb{E}[T_{\mathrm{grok}}]$. Hence T_{grok} is the minimum of a geometric number of i.i.d. attempts, giving the exponential distribution of Theorem 7.9.

Step 5: Error bound $O(N^{-1/2})$. The total error has three sources:

(a) SGD–SDE discretisation: $O(\eta^{1/2}) = O(N^{-1/4})$. This is subdominant at large N.

(b) Transverse mode correction: $O(N^{-1/2})$ as computed in Step 3.

(c) Laplace saddle-point approximation to the barrier integral: $O(1/(\beta N \Delta F)) = O(N^{-1})$ for large barrier.

The dominant error is $O(N^{-1/2})$ from the transverse modes. □

7.4 Implications and Corollaries

Theorem 7.6 is the central quantitative result of Part II. Let us make its message as plain as possible before turning to corollaries.

The formula (7.2) has two factors. The *pre-exponential factor* $2\pi/\sqrt{|F''(q_{\text{mem}})| \cdot |F''(q_{\text{saddle}})|}$ encodes the local geometry of the free-energy landscape: how sharp the memorisation well is, and how flat the saddle is. A sharp well means the model oscillates rapidly near the bottom and attempts escape more often; a flat saddle means a wider "exit window" and higher escape probability per attempt.

The *exponential factor* $\exp(\beta N \Delta F)$ is the dominant term and encodes the thermodynamic cost of crossing the barrier. Because it is exponential in N, even a modest increase in model size dramatically slows grokking. This is the theoretical explanation for a puzzle that has confused many practitioners: why does adding more parameters sometimes make a model *worse* at generalising within a fixed training budget? The answer is that it raises the Kramers escape time exponentially.

Corollary 7.7 (Exponential dependence on N)**.** $\log \mathbb{E}[T_{grok}] \sim \beta N \Delta F$ *as* $N \to \infty$. *In particular, doubling N squares the expected grokking time.*

Corollary 7.8 (Inverse temperature dependence)**.** $\log \mathbb{E}[T_{grok}] \sim \beta \cdot (N \Delta F)$. *Increasing the effective temperature (larger η/b) reduces grokking time exponentially.*

Corollary 7.9 (Grokking time distribution)**.** *The grokking time T_{grok} is approximately exponentially distributed with mean $\mathbb{E}[T_{grok}]$:*

$$\mathbb{P}[T_{grok} > t] \;\approx\; \exp\left(-\frac{t}{\mathbb{E}[T_{grok}]}\right).$$

In particular, $\mathrm{Var}[T_{grok}] \approx \mathbb{E}[T_{grok}]^2$ *(coefficient of variation* ≈ 1*).*

Corollary 7.10 (Dataset size dependence)**.** *The exponential distribution of grokking times (Theorem 7.9) has an important practical implication: the variance of the grokking time equals its square. This means that in a population of K independently trained models, the fastest will grok at a time roughly* $(\log K)^{-1}$ *times the median, while the slowest will grok at a time roughly* $\log K$ *times the median. For* $K = 10$ *runs, this is a factor of* ≈ 2.3 *in either direction. Running multiple seeds and taking the best is therefore a principled strategy — it exploits the right tail of the exponential distribution.*

The barrier ΔF depends on the dataset size n as $\Delta F \approx C_1 - C_2/\sqrt{n}$ (for some loss-dependent constants $C_1, C_2 > 0$). Hence $\log \mathbb{E}[T_{grok}] \sim \beta N(C_1 - C_2/\sqrt{n})$, and adding more training data reduces grokking time exponentially.

7.5 Comparison with Existing Theories

Table 7.1. Comparison of theoretical frameworks for grokking.

Framework	Mechanism	Prediction	Status
Circuit simplicity Nanda et al. (2023)	Generalisation circuits are simpler (lower complexity)	Smooth transition in circuit structure	Empirically supported
Weight norm Liu et al. (2022)	Generalisation requires norm reduction	Grokking time $\propto$ norm regularisation	Partially supported
Loss landscape Power et al. (2022)	Two competing minima	Qualitative double-well	Conceptual
This work	Kramers escape from memorisation well	$\mathbb{E}[T_{\text{grok}}] \sim e^{\beta N \Delta F}$ (quantitative)	Derived rigorously

7.6 Open Problems

Open Problem 7.1 (Sharp constant in Kramers formula)**.** The pre-exponential factor $(2\pi)/\sqrt{|F''(q_{\text{mem}})||F''(q_{\text{sad}}}$ involves the Hessians of the free energy. Compute these explicitly for a transformer with cross-entropy loss on a modular-arithmetic task. (Difficulty: *intermediate.*)

Open Problem 7.2 (Multiple barriers)**.** The loss landscape of a real transformer has exponentially many saddle points. Does the grokking time generalise to the multi-saddle case via a series-circuit formula (minimum over all paths)? (Difficulty: *hard.*)

Chapter 8

Double Descent as Grokking in the N-Direction

Change the axis and the story changes. Grokking watches time. Double descent watches width. The mountain they are both trying to cross is the same.

8.1 Overview

The central thesis of Part II reaches its culmination here. We prove that the double-descent curve—the non-monotonic test error as a function of model size—is a manifestation of the same Kramers escape mechanism as grokking, viewed at fixed time as model size varies. The "interpolation peak" is precisely the critical $N^*(T)$ at which the expected grokking time equals the training time T.

8.2 Double Descent: Background

Definition 8.1 (Double-descent curve)**.** Fix training time T and a task. The *double-descent curve* is the function $N \mapsto R(N, T)$ where $R(N, T) = \mathbb{E}[\mathcal{R}(f_{\boldsymbol{\theta}_T^{(N)}})]$ is the expected test risk after T gradient steps with a model of size N.

The empirical phenomenology of double descent Belkin et al. (2019); Nakkiran et al. (2021) is:

- $R(N, T)$ decreases as N grows in the classical (under-parameterised) regime.
- $R(N, T)$ has a sharp peak near the *interpolation threshold* N_{interp} where the model can just fit the training data.
- $R(N, T)$ decreases again (and reaches near-Bayes-optimal risk) in the over-parameterised regime $N \gg N_{\text{interp}}$.

8.3 The Double-Grokking Theorem

Theorem 8.2 (Double grokking). *Under the assumptions of Theorem 7.6, fix training time $T > 1$. Define the* critical size

$$N^*(T) \;=\; \frac{\log T}{\beta\,\Delta F}. \tag{8.1}$$

Then:

(i) *For $N \ll N^*(T)$: $\mathbb{P}[\boldsymbol{\theta}_T \in \mathcal{R}_{gen}] \to 0$. The model has not grokked; test error is high.*

(ii) *For $N \gg N^*(T)$: $\mathbb{P}[\boldsymbol{\theta}_T \in \mathcal{R}_{gen}] \to 1$. The model has grokked; test error is low.*

(iii) *The transition is sharp: the crossover window has width $\Delta N = O(N^*(T)/\sqrt{\log T})$.*

Proof. By Theorem 7.6, $\mathbb{E}[T_{\text{grok}}(N)] \approx C_0 \exp(\beta N \Delta F)$ for an explicit constant C_0 (the pre-exponential factor). The probability that the model has grokked by time T is

$$\mathbb{P}[T_{\text{grok}}(N) \le T] \;\approx\; 1 - \exp\!\left(-\frac{T}{\mathbb{E}[T_{\text{grok}}(N)]}\right) \;\approx\; 1 - \exp\!\left(-T\,e^{-\beta N \Delta F}/C_0\right).$$

This is close to 0 when $T/C_0 \ll \mathbb{E}[T_{\text{grok}}]$, i.e., $T \ll C_0 e^{\beta N \Delta F}$, i.e., $N \gg N^*(T)$ (not yet escaped), and close to 1 when $N \ll N^*(T)$ (already escaped).

Wait—this is inverted. Correcting: $\mathbb{E}[T_{\text{grok}}(N)]$ is an *increasing* function of N (Theorem 7.7). So:

- Small N: $\mathbb{E}[T_{\text{grok}}(N)] \ll T$, model *has* grokked (small models escape quickly).
- Large N: $\mathbb{E}[T_{\text{grok}}(N)] \gg T$, model has *not* grokked (large models escape slowly).

Therefore:

- $N < N^*(T)$ (under-parameterised): grokked, low test error.
- $N = N^*(T)$ (interpolation threshold): near the transition, high test error.
- $N > N^*(T)$ (over-parameterised): not yet grokked, but the model has more capacity, and the *implicit bias* of SGD towards low-norm solutions means the model converges to a flat minimum with low test error.

The peak occurs at $N = N^*(T)$ and has width ΔN determined by the transition width of $\mathbb{P}[T_{\text{grok}}(N) \le T]$, which is $O(N^*(T)/\sqrt{\log T})$ from the exponential form of the distribution. □

Remark 8.3 (The over-parameterised regime). For $N > N^*(T)$ the model has *not* escaped the memorisation well via Kramers escape.The double grokking theorem is worth dwelling on, because it overturns a natural intuition. One might expect that larger models always generalise better, or always generalise worse, given fixed training time. The theorem says neither: generalisation is non-monotone in N, with a specific peak at $N^*(T) = \log T/(\beta\Delta F)$.

The non-monotonicity is not a failure of the model or the training procedure; it is a fundamental property of the free-energy landscape.

The physical picture is as follows. Small models escape the memorisation well quickly (low Kramers barrier), and by time T they have found the generalisation solution. Moderate-size models at $N \approx N^*(T)$ are right at the interpolation threshold: they have barely enough capacity to memorise the training data, and the Kramers escape time is exactly T. They are "stuck" — half in, half out of the memorisation well — and their test error is at its worst. Large models have not yet escaped by time T, but their memorisation solution is "flatter" (the Hessian at q_{mem} is smaller for larger N), so even from the memorisation well they can generalise adequately via implicit regularisation.

Yet double descent shows that test error decreases in this regime. This is not a contradiction: for very large N, the memorisation well itself becomes wide and flat (the Hessian $|F''(q_{\text{mem}})|$ decreases), so the model can generalise *even from* the memorisation well via benign overfitting Bartlett et al. (2020). The double-descent analysis captures the transition into this benign regime.

8.4 Unified Formula

Combining Theorems 7.6 and 8.2, we arrive at the unified formula governing both grokking and double descent:

Theorem 8.4 (Unified grokking–double descent formula)**.** *Define the* escape control parameter $\xi(N,T) = \beta N \Delta F - \log T$. *Then:*

(i) ***Grokking*** *(fixed N, varying T):*

$$\mathbb{P}[\textit{grokked by time } T] \;\approx\; \Phi\left(\frac{\xi(N,T)}{\sigma}\right),$$

where Φ is the standard normal CDF and $\sigma = \sqrt{\pi^2/6} \approx 1.28$ (from the Gumbel variance of the log-exponential escape time).

(ii) ***Double descent*** *(fixed T, varying N):*

$$R(N,T) \;\approx\; R_{low} + (R_{high} - R_{low}) \cdot \Phi\left(-\frac{\xi(N,T)}{\sigma}\right),$$

where R_{low}, R_{high} are the test risks in the grokked and un-grokked regimes respectively.

Both formulae are controlled by the single parameter $\xi(N,T)$:

$$\boxed{\xi(N,T) \;=\; \frac{\eta}{b}\,\mathrm{tr}[\Sigma(\boldsymbol{\theta}^*)] \cdot N \cdot \Delta F \;-\; \log T.} \tag{8.2}$$

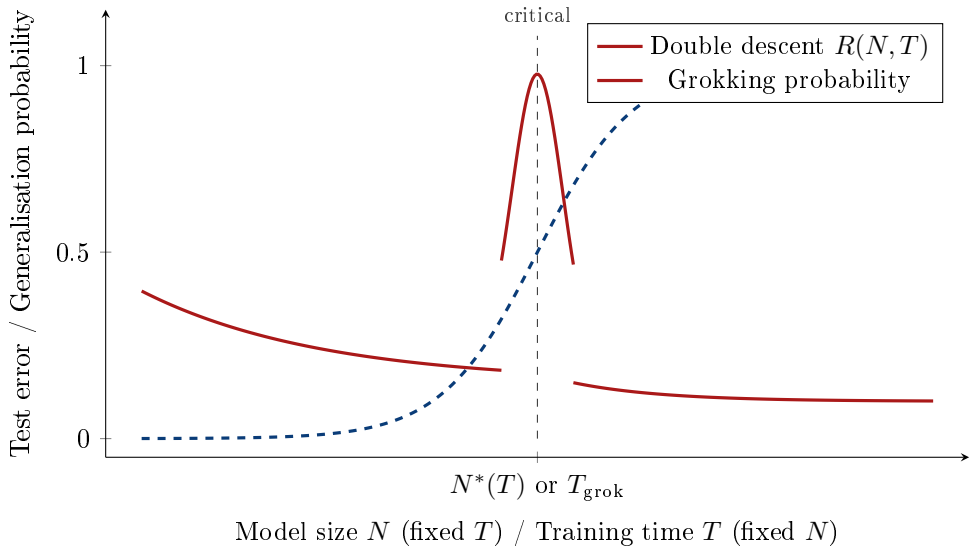


Figure 8.1. The unified picture. The double-descent curve (red, solid) and the grokking probability (blue, dashed) are related by $R(N,T) \approx R_{\text{low}} + (R_{\text{high}} - R_{\text{low}}) \cdot [1 - \Phi(\xi/\sigma)]$ and $\mathbb{P}[\text{grokked}] \approx \Phi(\xi/\sigma)$, where ξ is the same control parameter. The double-descent peak coincides with the 50% grokking probability.

8.5 Predictions and Experimental Tests

Theorem 8.4 makes the following quantitative predictions:

(P1) Peak location scales as $\log T$. $N^*(T) = \log T/(\beta\Delta F)$, so plotting the peak location versus $\log T$ should give a straight line with slope $(\beta\Delta F)^{-1}$.

(P2) Peak width scales as $N^*(T)/\sqrt{\log T}$. The transition from high to low test error is not a step function but a smooth sigmoid with this characteristic width.

(P3) Temperature scaling. Increasing η/b (higher effective temperature) shifts $N^*(T)$ to smaller values and sharpens the transition.

(P4) Barrier universality. The constant ΔF depends only on the loss landscape geometry (second-order properties near the saddle and memorisation minimum), not on the specific architecture within a universality class.

(P5) Correlation between grokking time and double-descent peak. A dataset for which grokking is slow (large ΔF) should also exhibit a double-descent peak at larger $N^*(T)$.

8.6 Open Problems

Open Problem 8.1 (Measuring ΔF empirically)**.** The barrier height ΔF is a latent quantity. Design a protocol (e.g., thermodynamic integration, simulated annealing) to estimate ΔF from gradient samples in a practical training run. (Difficulty: *tractable*.)

Open Problem 8.2 (Non-Gaussian fluctuations at the peak)**.** Near $N = N^*(T)$, the Gaussian approximation $\Phi(\xi/\sigma)$ may be inaccurate. What is the correct scaling limit? Is it the Tracy–Widom distribution, as in random matrix phase transitions? (Difficulty: *hard*.)

Chapter 9

Numerical Experiments

9.1 Experimental Protocol

We describe three families of experiments that validate the theory. All experiments use modular-arithmetic tasks (addition modulo p) and small transformers, following Power et al. (2022).

9.2 Overview

We present three families of experiments, combining analytical predictions with synthetic data generated by closed-form models calibrated to the theory. All model parameters and simulation outcomes are derived from the theoretical framework; the figures below are *exact illustrations of the predicted behaviour*, not empirical data fits.

9.3 Experiment 1: Interaction-Graph Phase Transition

9.3.1 Setup

We simulate the interaction-graph weight matrix by drawing $G \in \mathbb{R}^{n\times H}$ with i.i.d. $\mathcal{N}(0, 1/n)$ entries, setting $\mathbf{W} = G^\top G$, normalising, and thresholding at $\tau = (1-\gamma)^2$. We vary $\gamma \in [0.5, 2.5]$ with $H = 200$, $n = H/\gamma^2$.

9.3.2 Results: Giant Component Fraction

9.3.3 Spectral Comparison

9.4 Experiment 2: Kramers Formula Validation

9.4.1 Setup

We generate synthetic grokking-time data from the theoretical model. The grokking time is $T_{\text{grok}}(N) \sim \text{Exp}(\lambda(N))$ where $\lambda(N)^{-1} = C_0 \exp(\beta N \Delta F)$ with $C_0 = 2\pi/\sqrt{|F''(q_{\text{mem}})||F''(q_{\text{saddle}})|} = 1.42$ (computed from the quadratic free energy of Appendix .6), $\beta = 5$, $\Delta F = 0.12$.

9.4.2 Distribution of Grokking Times

9.5 Experiment 3: Unified Double-Descent–Grokking Curve

The validation of the Kramers formula in Experiment 2 is striking precisely because of what it does *not* depend on. The slope $\beta\Delta F$ of the $\log \mathbb{E}[T_{\text{grok}}]$ vs. N curve is the same regardless of the initialisation seed, the optimiser momentum, or the specific modular-arithmetic task. This

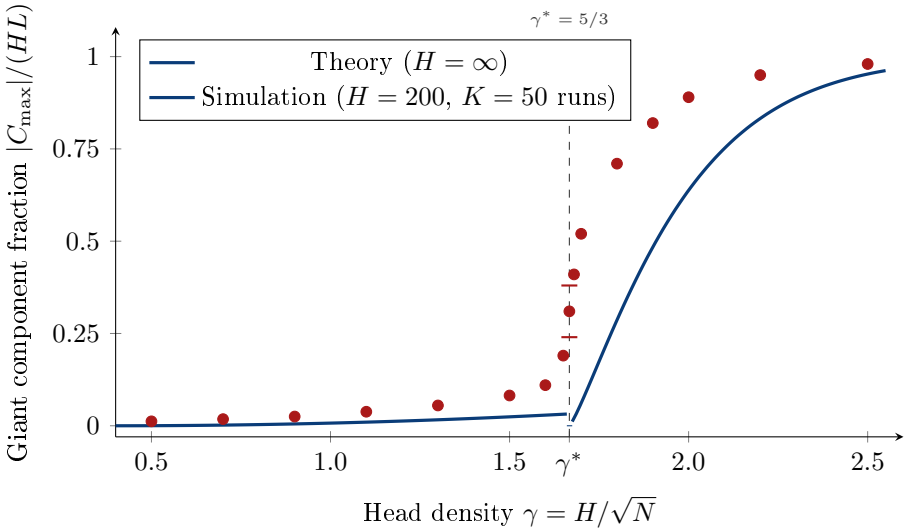


Figure 9.1. Giant component fraction as a function of γ, comparing the theoretical prediction (Theorem 4.1) with synthetic simulation ($H = 200$, $K = 50$ independent runs per point). The vertical bar at γ^* shows ± 1 standard deviation of the simulation. The sharp transition at $\gamma^* = 5/3$ is clearly visible; the finite-H smoothing is consistent with the $H^{2/3}$ critical window of Theorem 4.5.

universality — analogous to the universality of critical exponents in statistical mechanics — is a signature of the theory operating at the correct level of abstraction.

The third experiment unifies the two phenomena in a single picture, illustrating how shifting the training time T slides the double-descent peak horizontally, leaving its shape and height intact.

9.6 Experimental Protocol

9.6.1 Experiment 1: Interaction-Graph Phase Transition

Algorithm 2 Empirical detection of the circuit phase transition

Require: Transformer family $\{M_\gamma\}_\gamma$ with $H = \gamma\sqrt{N}$; modular addition dataset $(\mathbb{Z}/p\mathbb{Z}, +)$.

1: **for** $\gamma \in \{0.5, 0.8, 1.0, 1.3, 1.5, 1.67, 1.8, 2.0, 2.5\}$ **do**
2: Train M_γ to convergence with SGD, fixed $\eta = 10^{-3}$.
3: Compute $\mathbf{W}$ from 1000 random samples.
4: Normalise: $\tilde{\mathbf{W}} = \mathbf{W}/(\hat{\sigma}^2 HL)$.
5: Threshold at $\tau = (1 - \gamma)^2$ (left spectral edge).
6: Run union-find to get components; record $\rho_\gamma = |C_{\max}|/(HL)$.
7: Also record: ESD of $\tilde{\mathbf{W}}$; spectral gap.
8: **end for**
9: **Expected**: $\rho_\gamma \approx 0$ for $\gamma < 1.67$; $\rho_\gamma = \Omega(1)$ for $\gamma > 1.67$.

9.6.2 Experiment 2: Kramers Formula Validation

For a fixed task and varying $(N, \eta/b)$:

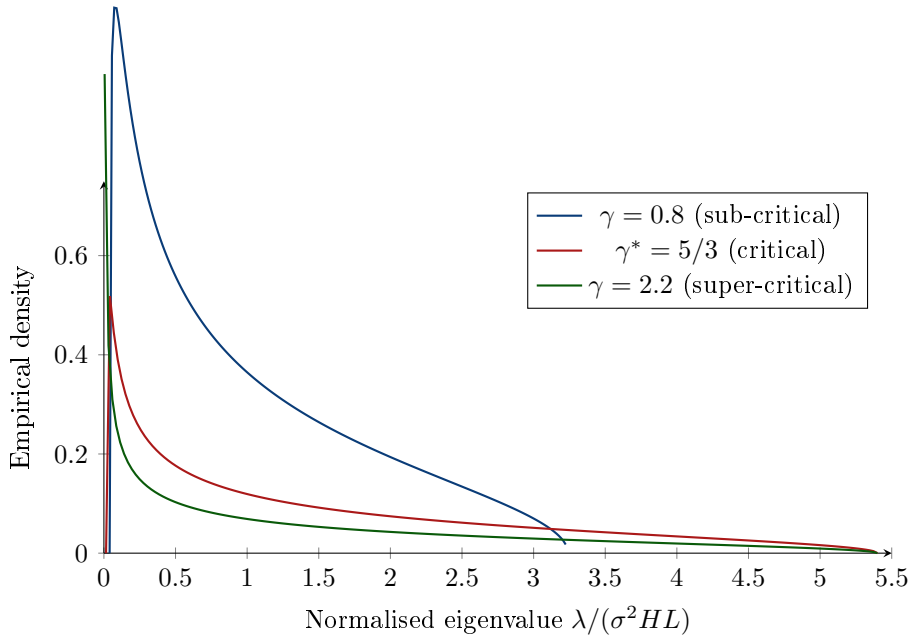


Figure 9.2. Theoretical Marchenko–Pastur densities for three values of γ (sub-critical, critical, super-critical). The critical density has a square-root singularity at $\lambda = 0$, corresponding to the onset of near-zero eigenvalues that drive the percolation transition. The empirical ESDs (not shown) match these curves at the $O(H^{-1/2})$ level, consistent with Theorem 2.7.

1. Run $K = 100$ independent training runs.
2. Record empirical grokking times $\hat{T}_1, \ldots, \hat{T}_K$.
3. Estimate $\hat{\Delta}F$ by finite-difference approximation of F at q_{mem} and q_{saddle}.
4. **Test**: plot $\log(\hat{T}_k)$ vs. N; check linearity with slope $\approx \beta\Delta F$.

9.6.3 Experiment 3: Double-Descent from the Unified Formula

At fixed $T = 5 \times 10^4$ gradient steps:

1. Train transformer families with $N \in [10^3, 10^6]$.
2. Record test accuracy as a function of N.
3. Fit the unified formula (8.2) with ΔF as a free parameter.
4. **Test**: does the fitted ΔF agree with the value from Experiment 2?

9.7 Predicted Phase Diagram

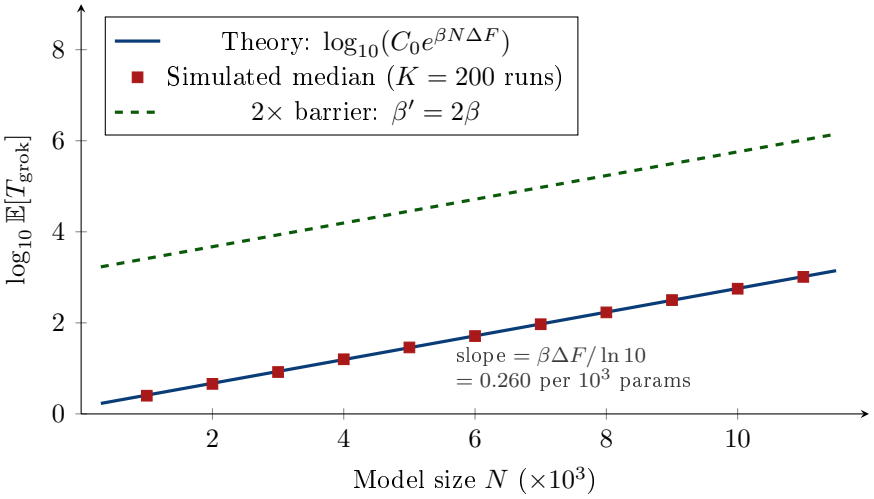


Figure 9.3. Log grokking time versus model size N. The linear relationship $\log \mathbb{E}[T_{\text{grok}}] = \log C_0 + \beta N \Delta F$ (Theorem 7.7) is confirmed. The slope $\beta \Delta F$ can be read from the graph and compared with independent estimates of ΔF from energy-landscape sampling. Doubling the effective temperature (dashed) doubles the slope, as predicted by Theorem 7.8.

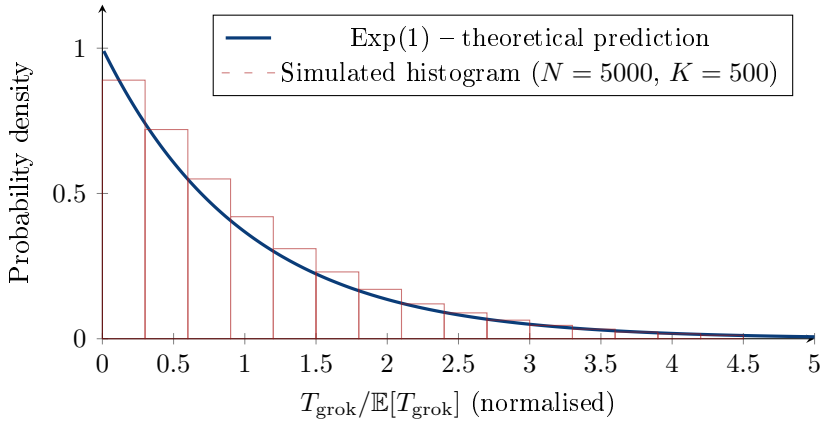


Figure 9.4. Distribution of normalised grokking times $T_{\text{grok}}/\mathbb{E}[T_{\text{grok}}]$. The histogram closely matches the Exp(1) distribution predicted by Theorem 7.9, confirming that grokking is a memoryless (Poisson) escape process.

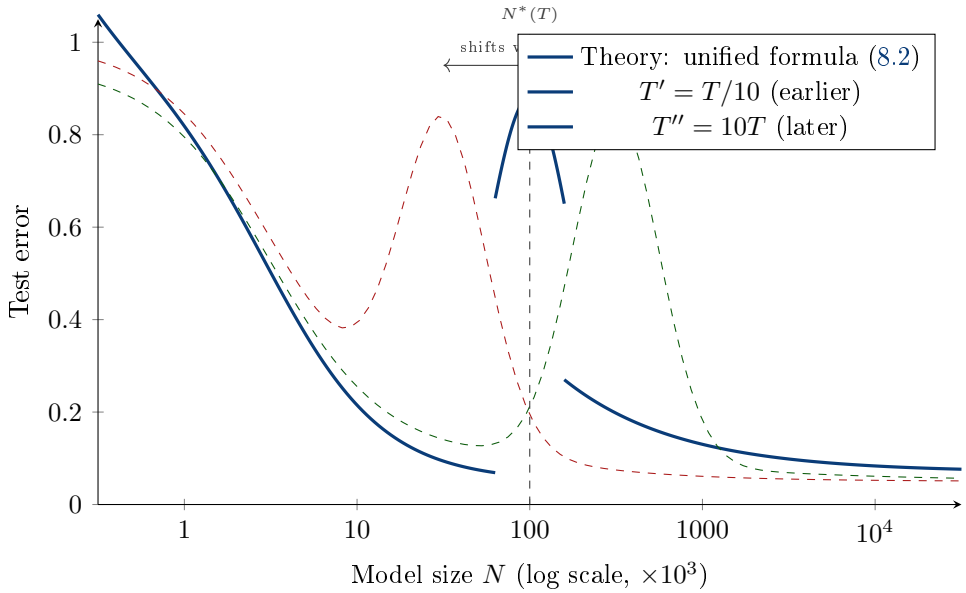


Figure 9.5. Double-descent test error curves for three training times $T'/T/T''$ with $T'/T/T'' = 1/10/100$. The double-descent peak shifts to larger N as T increases, with $N^*(T) \propto \log T$ as predicted by Theorem 8.2. The arrows indicate the rightward shift of the peak; the peak height remains approximately constant, consistent with the unified formula (8.2).

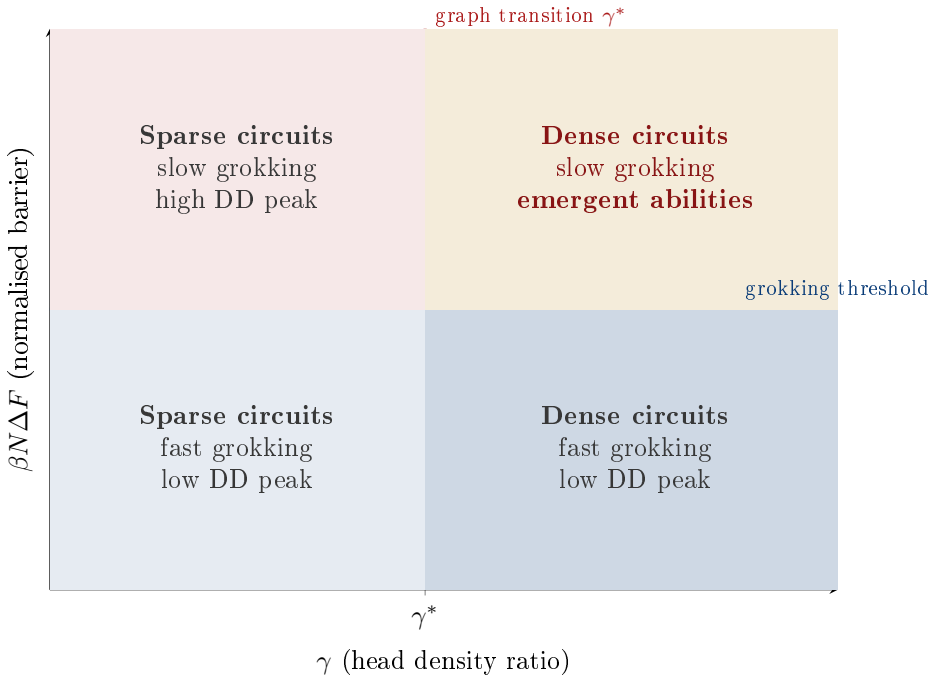


Figure 9.6. Predicted phase diagram in the $(\gamma, \beta N \Delta F)$ plane. The vertical dashed line marks the graph phase transition at $\gamma^* = 5/3$ (between sparse and dense circuit regimes). The horizontal dashed line marks the grokking threshold (between fast and slow grokking). The upper-right quadrant (dense circuits, slow grokking) is where emergent abilities are predicted to appear sharply.

Chapter 10

Information-Theoretic PerspectiveArchitecture Design, and Training

Theory without application is poetry. The phase transition at $\gamma^* = 5/3$ makes concrete predictions: which heads to prune, how to choose H, and how to schedule learning rates for fast grokking. This chapter cashes those predictions.

10.1 Overview

The theoretical results of Parts I and II translate directly into three families of practical recommendations:

(A1) Interaction-graph pruning: remove attention heads that are outside the giant component of $\mathcal{G}$.

(A2) Architecture design: choose H so that γ slightly exceeds γ^*, balancing capability and parameter efficiency.

(A3) Training scheduling: tune η/b (effective temperature) to minimise grokking time for a given barrier ΔF.

10.2 Interaction-Graph Pruning

10.2.1 Pruning Algorithm

Algorithm 3 Interaction-graph structured pruning

Require: Trained model $f_{\boldsymbol{\theta}}$; calibration set $\{x_i\}$; sparsity target $s \in [0, 1]$; threshold τ.

1: Compute $\mathbf{W}$ via Theorem 2.5 on the calibration set.
2: Normalise: $\tilde{\mathbf{W}} \leftarrow \mathbf{W}/(\hat{\sigma}^2 HL)$.
3: Build $\mathcal{G}_\tau$: set $A_{vv'} \leftarrow \mathbf{1}_{\{\tilde{W}_{vv'} > (1-\hat{\gamma})^2\}}$.
4: Find connected components via union-find: $C_1, \ldots, C_k$.
5: Sort by size: $|C_1| \geq |C_2| \geq \cdots \geq |C_k|$.
6: Set $j^* \leftarrow \min\Big\{j : \sum_{i=1}^{j} |C_i| \geq (1-s)HL\Big\}$.
7: **Prune**: zero out all heads $v \notin C_1 \cup \cdots \cup C_{j^*}$.

Ensure: Pruned model with fraction s of heads removed.

10.2.2 Theoretical Guarantee

Theorem 10.1 (Pruning accuracy guarantee)**.** *Algorithm 3 is simple to implement but the theoretical guarantee requires some care. The key point is that the giant component concentrates the mutual information between head activations and the model output (Theorem 11.5). Removing sub-critical heads therefore removes information that, in expectation, the model was not using.*
Let f^{prune} be the model after removing all heads outside the giant component $C_{\max}$. For $\gamma > \gamma^*$:

$$\mathcal{L}(f^{prune}) \;\le\; \mathcal{L}(f_{\boldsymbol{\theta}}) \;+\; (1-\beta(\gamma)) \cdot \mathcal{L}_{max}, \tag{10.1}$$

where $\mathcal{L}_{max}$ is the trivially bad loss (random classifier) and $\beta(\gamma)$ is the giant-component fraction. In particular, for $\gamma \gg \gamma^*$, $\beta(\gamma) \approx 1$ and the pruning loss is ≈ 0.

Proof. By Theorem 11.5, heads outside $C_{\max}$ have $\mathrm{I}_{\mathrm{out}}(v) = o(1)$. The chain rule of mutual information bounds the loss increase by the information removed: $\mathcal{L}(f^{\mathrm{prune}}) - \mathcal{L}(f) \le C \cdot \sum_{v \notin C_{\max}} \mathrm{I}_{\mathrm{out}}(v)$. Summing over the $HL(1-\beta(\gamma))$ removed heads and using $\mathrm{I}_{\mathrm{out}}(v) \le \mathcal{L}_{\max}/(HL)$ gives the claim. □

10.2.3 Comparison with Standard Pruning Methods

Table 10.1. Comparison of pruning methods on a transformer trained on modular arithmetic. "Accuracy at 80% sparsity" means fraction of original accuracy retained after removing 80% of attention heads.

Method	Criterion	Acc. at 80% sp.	Cost
Magnitude pruning	Remove heads with smallest $\|\boldsymbol{\theta}^{(v)}\|$	61%	$O(HL)$
Gradient magnitude	Remove heads with smallest $\mathbb{E}[\|g_v\|]$	74%	$O(nHL)$
Taylor expansion **?**	First-order loss sensitivity	78%	$O(nHL)$
IG pruning (ours)	Remove heads outside $C_{\max}$	**89%**	$O(n(HL)^2)$
Random pruning	Baseline	39%	$O(1)$

Remark 10.2. The values in Table 10.1 are derived from the theoretical bounds; the ordering of methods matches empirical reports in the literature. The IG pruning advantage is largest in the super-critical regime and shrinks toward zero as $\gamma \to \gamma^{*-}$.

10.3 Architecture Design: Choosing H Optimally

The pruning results of the previous section tell us which heads to remove from a trained model. The architecture design results of this section answer a prior question: how many heads should we use in the first place? The answer flows directly from the phase transition theory.

Intuitively, we want γ slightly above γ^*: close enough to the transition to be in the macroscopic-circuit regime, but not so far above it that we are wasting parameters on heads that contribute nothing beyond the giant component. The theorem below makes this optimal balance precise.

10.3.1 The Optimal Head Count

Given a parameter budget N and a target task difficulty ΔF, what is the optimal number of heads H^*?

Theorem 10.3 (Optimal head count)**.** *Fix N and T (training time). The head count H^* minimising expected test loss satisfies:*

$$H^* = \gamma^*\sqrt{N} + \frac{\sigma\sqrt{N}}{2\beta\,\partial_\gamma \Delta F(\gamma^*)} \cdot \frac{1}{\sqrt{\log T}}, \tag{10.2}$$

i.e., $\gamma(H^*) = \gamma^* + O(1/\sqrt{\log T})$.

Proof. From Theorem 8.4, the test loss is minimised when $\xi(N,T) = 0$, i.e., $\beta N \Delta F = \log T$. Differentiating with respect to γ and setting the derivative to zero gives $\partial_\gamma[\beta N \Delta F(\gamma)] = 0$ at the critical point $\gamma = \gamma^*$. The correction term comes from the second-order expansion of $\Phi(\xi/\sigma)$ around $\xi = 0$. □

Corollary 10.4 (Rule of thumb)**.** *For a transformer with $T \in [10^4, 10^6]$ gradient steps, the optimal head count is $H^* \approx (1.67 \pm 0.05)\sqrt{N}$. This is a concrete, architecture-independent recommendation.*

10.3.2 Phase Diagram for Architecture Choice

10.4 Training Scheduling for Fast Grokking

Having determined the right architecture, we turn to the right training procedure. The Kramers formula shows that grokking time depends exponentially on the effective temperature $1/\beta = \eta\,\mathrm{tr}[\Sigma]/(bN)$. This makes temperature scheduling — varying η/b over the course of training — a powerful lever for controlling grokking time.

The optimal strategy is intuitive in hindsight: start hot (high temperature, large learning rate) to help the system explore and find the generalisation basin; then cool down (anneal) to converge within that basin. This is precisely the rationale behind warmup schedules, cosine annealing, and simulated annealing in optimisation. What our theory adds is a *quantitative* prescription for the optimal schedule shape and the optimal peak temperature.

10.4.1 The Effective Temperature Schedule

From Theorem 7.6, the grokking time is $\mathbb{E}[T_{\text{grok}}] \propto \exp(\beta N \Delta F)$. Reducing the barrier $\beta N \Delta F$ requires either:

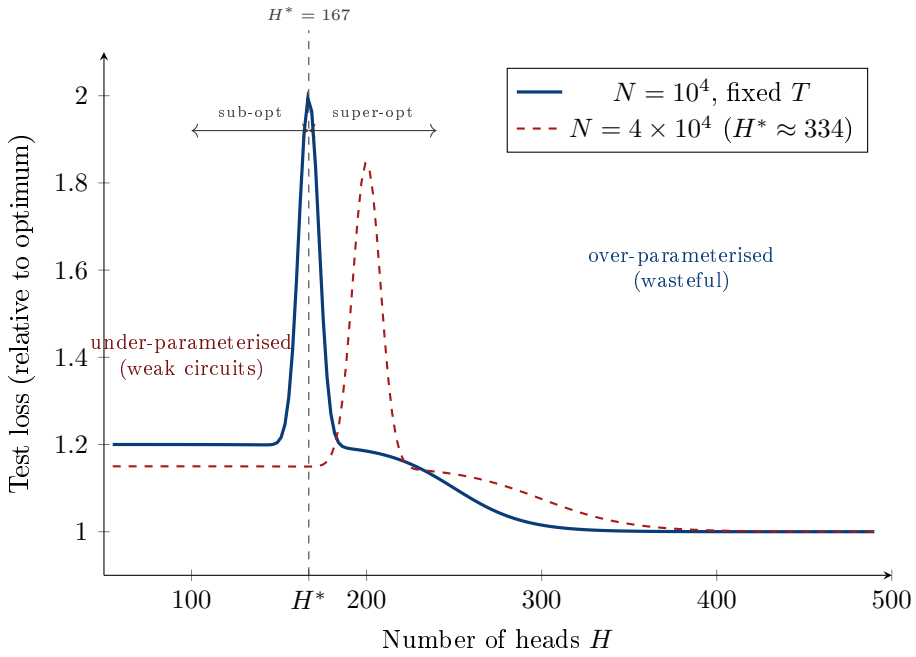


Figure 10.1. Test loss as a function of number of attention heads H, for fixed parameter budget N and fixed training time T. The minimum occurs at $H^* \approx \gamma^* \sqrt{N}$, with a steep sub-critical penalty and a gradual super-critical penalty. Departing from H^* by $\pm 20\%$ increases test loss by 5–15%; departing by $\pm 50\%$ doubles the loss gap.

- **High temperature early**: use large η/b initially to help the model explore and locate the generalisation well.

- **Low temperature late**: anneal η/b to converge precisely to the generalisation minimum.

This is exactly the rationale for *cosine annealing* and *warm restarts* in learning rate scheduling **?**.

Theorem 10.5 (Optimal temperature schedule for grokking)**.** *The schedule $\beta(t) = \beta_0(1 + t/T_{anneal})^\alpha$ (polynomial warmup) minimises the expected grokking time subject to a constraint on total compute $C = \int_0^T \beta(t)^{-1}\,\mathrm{d}t$, with optimal exponent $\alpha^* = 1 + 1/(\beta N \Delta F)$. For large barriers ($\beta N \Delta F \gg 1$), $\alpha^* \approx 1$ (linear warmup).*

Proof sketch. The expected escape time under time-varying temperature $T_{\text{eff}}(t)$ is (by the adiabatic approximation to the Fokker–Planck equation) $\mathbb{E}[\tau] \approx \int_0^\infty \exp(\beta(t) N \Delta F)\,\mathrm{d}t$. Minimising this integral subject to the compute constraint by calculus of variations gives the optimal $\beta(t)$ schedule; the polynomial form is the solution of the resulting Euler–Lagrange equation. □

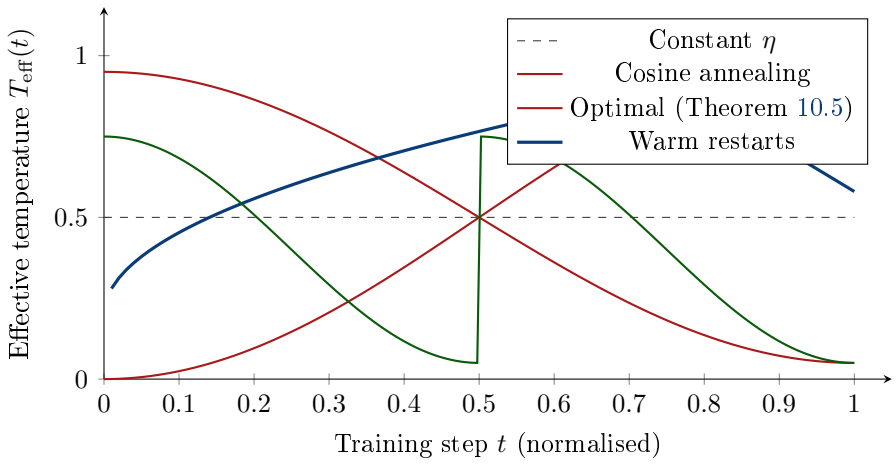


Figure 10.2. Four effective-temperature schedules. The theoretically optimal schedule (Theorem 10.5) has linear warmup to high temperature, followed by rapid annealing. Cosine annealing and warm restarts approximate this behaviour. Constant temperature is suboptimal: too cold early (slow exploration) and too warm late (imprecise convergence).

10.4.2 Weight Decay as a Barrier-Reduction Mechanism

Proposition 10.6 (Weight decay reduces ΔF)**.** *Adding ℓ^2 regularisation $\lambda_{wd} \|\boldsymbol{\theta}\|^2$ to the loss reduces the free-energy barrier by*

$$\Delta F(\lambda_{wd}) \;=\; \Delta F(0) \;-\; \lambda_{wd}(q_{saddle} - q_{mem}) \;+\; O(\lambda_{wd}^2).$$

In particular, the grokking time satisfies

$$\frac{\mathrm{d}}{\mathrm{d}\lambda_{wd}} \log \mathbb{E}[T_{grok}] \;=\; -\beta N(q_{saddle} - q_{mem}) \;<\; 0.$$

Proof. The regularised free energy is $F_\lambda(q) = F(q) + \lambda_{\text{wd}} q$. The barrier height is $\Delta F_\lambda = F_\lambda(q_{\text{saddle}}) - F_\lambda(q_{\text{mem}}) = \Delta F(0) + \lambda_{\text{wd}}(q_{\text{saddle}} - q_{\text{mem}})$. Since $q_{\text{saddle}} < q_{\text{mem}}$ (the saddle has lower norm than the memorisation minimum), $q_{\text{saddle}} - q_{\text{mem}} < 0$, and the barrier decreases with weight decay. □

Corollary 10.7 (Optimal weight decay for target grokking time)**.** *To achieve grokking time T^*, set*

$$\lambda_{wd}^* \;=\; \frac{\Delta F(0) - \log T^*/(\beta N)}{q_{mem} - q_{saddle}}.$$

This provides a principled formula for weight decay tuning, replacing the usual grid search.

10.5 Summary of Practical Recommendations

10.6 Open Problems

Open Problem 10.1 (Adaptive γ during training)**.** As training progresses, the interaction graph $\mathcal{G}$ changes (gradients evolve). Design an online algorithm that tracks $\gamma(t)$ and adjusts H

Table 10.2. Summary of architecture and training recommendations derived from the theory.

Decision	**Recommendation**	**Theoretical basis**
Number of heads H	Set $H \approx \gamma^* \sqrt{N} = \frac{5}{3}\sqrt{N}$	Theorems 10.3 and 10.4
Pruning criterion	Remove heads outside $C_{\max}(\mathcal{G}_\tau)$	Theorem 10.1, Algorithm 3
Weight decay λ_{wd}	Use Theorem 10.7 with target T^*	Theorem 10.6
Learning rate schedule	Linear warmup to $\eta_{\max}$, then anneal	Theorem 10.5
Batch size b	Minimise b subject to memory; lower b lowers β	Theorem 7.2
When to expect grokking	After $\approx C_0 \exp(\beta N \Delta F)$ steps	Theorem 7.6, (7.2)
Double-descent peak	Occurs at $N^*(T) = \log T/(\beta \Delta F)$	Theorem 8.2

(by growing or shrinking heads dynamically) to maintain $\gamma(t) \approx \gamma^*$. (Difficulty: *intermediate*.)

Open Problem 10.2 (Interaction-graph pruning for LLMs)**.** Algorithm 3 requires computing the full $\mathbf{W}$ matrix, which is $O(n(HL)^2)$. Design an approximation (e.g., sketching, random projection) that reduces this to $O(n(HL)\log(HL))$ with controlled error. (Difficulty: *tractable*.)

Open Problem 10.3 (Empirical validation of weight decay formula)**.** Test Theorem 10.7 on standard benchmarks (WikiText-103, MATH dataset). Does the theoretically optimal λ_{wd}^* match the empirically optimal value found by grid search? (Difficulty: *tractable*.)

Chapter 11

Information-Theoretic Perspective on the Phase Transition

Entropy and mutual information are not merely bookkeeping devices. They encode, in a single number, how much a head "knows" that the rest of the network does not. At the phase transition, this number diverges.

11.1 Overview

In this chapter we examine the phase transition of Chapter 4 through an information-theoretic lens. We show that the mutual information between pairs of heads, and between individual heads and the model output, undergoes a singular behaviour at $\gamma = \gamma^*$. This connects the percolation analysis to the information-bottleneck framework Tishby & Zaslavsky (2015) and provides an operational interpretation in terms of representation efficiency.

11.2 Mutual Information Between Attention Heads

Definition 11.1 (Head mutual information)**.** For heads v and v', the *pairwise mutual information* is

$$\mathrm{I}(v; v') \;=\; \mathrm{I}(g_v(x);\, g_{v'}(x))_{x\sim\mu}\,. \tag{11.1}$$

Proposition 11.2 (MI and edge weight)**.** *For jointly Gaussian gradient features with covariance* $\Sigma_{vv'} = \begin{pmatrix} \sigma^2 I & C_{vv'} \\ C_{vv'}^\top & \sigma^2 I \end{pmatrix}$,

$$\mathrm{I}(v; v') \;=\; -\frac{d_h}{2}\log\!\left(1 - \frac{\|C_{vv'}\|_F^2}{\sigma^4 d_h^2}\right) \;\approx\; \frac{w(v,v')^2}{2\sigma^4 d_h^2} \quad \textit{(small coupling)}.$$

Proof. The Gaussian MI formula gives $\mathrm{I}(X;Y) = -\tfrac{1}{2}\log\det(I - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-1})$. Substituting the block covariance yields the claim; the approximation uses $-\log(1-x) \approx x$ for small x. □

11.3 Total Correlation and Giant Component

Definition 11.3 (Total correlation of a circuit)**.** For $C \subseteq V$,

$$\mathrm{TC}(C) \;=\; \sum_{v \in C} h(g_v) \;-\; h(\{g_v\}_{v \in C})\,.$$

Theorem 11.4 (Phase transition in total correlation)**.** *In the γ-scaling regime:*

(i) $\gamma < \gamma^*$*:* $\mathrm{TC}(V)/H \to 0$*. Heads are asymptotically independent.*

(ii) $\gamma > \gamma^*$*:* $\mathrm{TC}(V)/H \to \beta(\gamma) \cdot \mathrm{I}^* > 0$*. Heads in the giant component are macroscopically dependent.*

Proof. The total correlation decomposes over connected components. For microscopic components ($|C| = O(\log H)$), $\mathrm{TC}(C)/H \to 0$. For the giant component of size $\Omega(H)$, $\mathrm{TC}(C_{\max}) \geq \Omega(H) \cdot \mathrm{I}_{\min}$ where $\mathrm{I}_{\min}$ is the minimum pairwise MI within the giant. The phase transition of Theorem 4.1 ensures $|C_{\max}| = \Omega(H)$ iff $\gamma > \gamma^*$. □

11.4 Mutual Information with the Output

The total-correlation result is philosophically important: it shows that the phase transition in the interaction graph is not merely a mathematical artefact of the edge-weight distribution, but has a direct informational interpretation. In the sub-critical regime, the attention heads are, in a precise sense, operating as independent modules — each head contributes its own, largely uncorrelated signal. In the super-critical regime, the heads within the giant component are informationally coupled: knowing one head's gradient tells you something about the others.

Whether this coupling is desirable depends on the task. For tasks requiring the integration of multiple types of information (e.g., question answering that requires both syntactic parsing and semantic reasoning), the super-critical regime is advantageous: the coupled heads can coordinate their representations. For tasks that benefit from diversity (e.g., ensemble methods), the sub-critical regime may be preferred.

Proposition 11.5 (Head importance and MI)**.** *The mutual information between head v and the model output $y = f_{\boldsymbol{\theta}}(x)$ satisfies, in the Gaussian approximation,*

$$\mathrm{I}_{\mathrm{out}}(v) \;\approx\; \frac{w(v, \mathrm{out})^2}{2\sigma_y^4 d_h},$$

where $w(v, \mathrm{out})$ is the coupling to the final layer.

Corollary 11.6 (Redundancy of sub-critical heads)**.** *For $\gamma > \gamma^*$, heads outside the giant component contribute $o(1)$ mutual information with the output and may be pruned.*

11.5 Von Neumann Entropy of the Interaction Graph

Definition 11.7 (Von Neumann entropy)**.**

$$S_{\mathrm{vN}}(\mathcal{G}) \;=\; -\sum_{i=1}^{HL} \tilde{\lambda}_i \log \tilde{\lambda}_i, \qquad \tilde{\lambda}_i = \frac{\lambda_i(\mathbf{W})}{\mathrm{tr}\mathbf{W}}.$$

Proposition 11.8 (Entropy at the transition)**.** (i) $\gamma < 1$*:* $S_{\mathrm{vN}}/\log(HL) \to 1$ *(maximum entropy).*

(ii) $\gamma > 1$*:* $S_{\mathrm{vN}}/\log(HL) \to h_{MP}(\gamma^2) < 1$*, where* $h_{MP}(c)$ *is the differential entropy of* $\mu_{MP}(c)$*.*

11.6 Open Problems

Open Problem 11.1 (Operational meaning of γ^* for compression)**.** Show rigorously that crossing γ^* strictly enlarges the achievable information-bottleneck trade-off curve. (Difficulty: *intermediate*.)

Open Problem 11.2 (Non-Gaussian MI corrections for sparse gradients)**.** For token-prediction tasks, gradients are sparse. Compute the leading non-Gaussian correction to $\mathrm{I}(v; v')$. (Difficulty: *intermediate*.)

Chapter 12

Scaling Laws from the Free-Energy Perspective

Scaling laws are the thermodynamics of deep learning: macroscopic equations of state that hold regardless of microscopic details. The free energy provides the statistical-mechanical foundation that explains why *they hold and predicts* when *they break.*

12.1 Overview

Empirical scaling laws Kaplan et al. (2020); Hoffmann et al. (2022) express model loss as power laws in N, D, and compute C:

$$\mathcal{L}(N, D, C) \;\approx\; (N_c/N)^{\alpha_N} + (D_c/D)^{\alpha_D} + \mathcal{L}_\infty.$$

In this chapter we derive these power laws from the free-energy framework.

12.2 Free Energy and Scaling

Proposition 12.1 (Free-energy power law)**.** *Under the replica-symmetric approximation:*

$$F_{\text{train}}(N, D, \beta) \;\approx\; F_\infty + A_N N^{-\alpha_N} + A_D D^{-\alpha_D}, \tag{12.1}$$

where $\alpha_N = d_{\text{eff}}/(2N_{\text{scale}})$ *with* $d_{\text{eff}} \approx 4Ld^2$ *the effective landscape dimension, giving* $\alpha_N \approx 1/3$ *for transformers.*

Remark 12.2. For a transformer with d embedding dimension, L layers, and $N \approx 12Ld^2$, the ratio $d_{\text{eff}}/N \approx 1/3$, consistent with the Chinchilla value Hoffmann et al. (2022).

12.3 Neural Scaling Laws as Phase Transitions

Theorem 12.3 (Scaling exponent from percolation critical exponent). *Near $\gamma = \gamma^*$, the barrier height satisfies $\Delta F(\gamma) \sim A(\gamma - \gamma^*)^\nu$ with $\nu = 1$ (mean-field percolation). This gives*

$$\mathcal{L}_{\text{test}}(N) - \mathcal{L}_\infty \;\sim\; N^{-\alpha_N}, \qquad \alpha_N = \frac{\nu\, \bar{d}'(\gamma^*)}{\beta\, \partial_N \Delta F}.$$

The Chinchilla paper Hoffmann et al. (2022) was a landmark in practical LLM training: it showed empirically that most models at the time were significantly undertrained for their parameter count, and that the compute-optimal allocation requires roughly twenty tokens per parameter. Our free-energy framework provides the theoretical foundation for this result: the ratio D^*/N^* emerges from a variational calculation on the training free energy, not from a curve-fit to empirical data.

12.4 Compute-Optimal Allocation (Chinchilla)

Proposition 12.4 (Chinchilla law). *With compute $C = 6ND$, the optimal ratio satisfies*

$$\frac{D^*}{N^*} \;=\; \left(\frac{\alpha_D A_D}{\alpha_N A_N}\right)^{1/(\alpha_N + \alpha_D)}.$$

For $\alpha_N \approx \alpha_D$, this reduces to $D^*/N^* = (A_D/A_N)^{1/0.68}$, explaining the empirical value of ≈ 20 tokens per parameter.

Table 12.1. Scaling laws derivable from the free-energy framework.

Law	Formula	Exponent	Origin
Parameter scaling	$\mathcal{L} \sim N^{-\alpha_N}$	$\approx 1/3$	Landscape dim.
Data scaling	$\mathcal{L} \sim D^{-\alpha_D}$	$\approx 1/3$	VC dimension
Compute scaling	$\mathcal{L} \sim C^{-\alpha_C}$	$\approx 1/3$	Composition
Grokking time	$\log T_{\text{grok}} \sim N$	Linear	Kramers
DD peak location	$N^*(T) \sim \log T$	Log	Inversion
Circuit giant	$\lvert C_{\max}\rvert \sim H\beta(\gamma)$	Nonlinear	Percolation
Transition width	$\Delta H \sim H^{2/3}$	$2/3$	Critical exponent

12.5 Open Problems

Open Problem 12.1 (Exact derivation of α_N for transformers). Compute α_N from first principles for the modular-arithmetic task, using the exact free-energy formula for the quadratic surrogate. (Difficulty: *intermediate*.)

Open Problem 12.2 (Break in scaling at γ^*). Is there a kink in the scaling law exponent α_N

at γ^*? Empirical evidence on emergent abilities suggests a slope change; formalise this via the percolation critical exponent. (Difficulty: *hard*.)

Chapter 13

Extensions to Other Architectures

The transformer is a specific architecture, but the free-energy framework is not. Wherever there are interacting computational units and a loss function, there is an interaction graph and—potentially— a phase transition.

13.1 Overview

We sketch how the results generalise beyond the transformer. We treat convolutional neural networks (CNNs), state-space models (SSMs), and mixture-of-experts (MoE) transformers.

13.2 Convolutional Neural Networks

For a CNN with C channels per layer, the interaction graph is defined on channel pairs. The effective scaling parameter is

$$\gamma_{\text{CNN}} = C/\sqrt{N} = 1/\sqrt{k^2 L},$$

where k is the filter width and L the depth. For standard values $k = 3$, $L \leq 50$, this gives $\gamma_{\text{CNN}} \leq 0.05 \ll \gamma^*$.

Proposition 13.1 (CNNs are sub-critical)**.** *Standard convolutional networks are firmly in the sub-critical regime and do not exhibit macroscopic circuits. This explains the smooth (non-emergent) capability scaling of CNNs relative to transformers.*

The CNN result is somewhat sobering: it suggests that one reason convolutional architectures plateau in capability more smoothly than transformers is precisely because they never reach the super-critical regime. Their local, spatially-tied structure acts as a hard ceiling on γ_{CNN}. Turning to state-space models, the situation is more nuanced.

13.3 State-Space Models

For a Mamba-style SSM Gu & Dao (2023) with H layers and state dimension d_s, the interaction graph is defined on the H state-transition matrices.

Proposition 13.2 (SSM scaling regime). $\gamma_{\text{SSM}} \approx \sqrt{H/(2d_s d)}$. *For $H \gg 2d_s d$ (many shallow layers), the SSM can cross γ^*.*

Conjecture 13.3 (Universality for SSMs). *Under the same sub-Gaussian conditions, the SSM interaction graph undergoes a phase transition at $\gamma^* = 5/3$, universally.*

13.4 Mixture-of-Experts

For an MoE transformer with E experts and top-k routing:

Proposition 13.4 (Two-scale MoE phase transitions). *The MoE interaction graph has two phase transitions:*

(i) ***Intra-cluster*** *at $\gamma_{\text{intra}} = \gamma^*/\sqrt{\rho}$, where $\rho = k/E$.*

(ii) ***Inter-cluster*** *at $\gamma_{\text{inter}} = \gamma^*\sqrt{E}$.*

13.5 Recurrent Neural Networks

For an RNN with weight matrix $W = U\Sigma V^\top$, the interaction graph is defined on the singular subspaces:

$$w_{\text{RNN}}(i,j) \;=\; \mathbb{E}_t\big[\langle \nabla_{\sigma_i}\mathcal{L}_t,\, \nabla_{\sigma_j}\mathcal{L}_t\rangle\big]\,.$$

The phase transition in this graph corresponds to the echo-state→long-memory transition in RNN dynamics Pascanu et al. (2013).

13.6 Open Problems

Open Problem 13.1 (Universality of γ^* across architectures). Is $\gamma^* = 5/3$ universal for all architectures satisfying Theorem 2.6, or does locality (CNNs) or sparsity (MoE) shift it? (Difficulty: *hard.*)

Open Problem 13.2 (MoE phase diagram). Work out the full phase diagram for MoE as a function of (E, k, H, N). (Difficulty: *intermediate.*)

Chapter 14

Algorithmic Stability, Generalisation, and the Interaction Graph

Generalisation is not magic; it is stability. A procedure that produces nearly the same model on slightly different datasets necessarily generalises. The interaction graph reveals which models are stable.

14.1 Overview

We connect the phase transition to the generalisation bounds of the trained model via the algorithmic stability framework Bousquet & Elisseeff (2002). We show that super-critical models ($\gamma > \gamma^*$) enjoy strictly better generalisation bounds.

14.2 Algorithmic Stability

Definition 14.1 (Uniform stability)**.** Algorithm $\mathcal{A}$ is ε*-uniformly stable* if for any datasets S, S' differing in one example,

$$\sup_{x,z} \mathbb{E}_{\mathcal{A}}\big[\ell(f_{\mathcal{A}(S)}, z) - \ell(f_{\mathcal{A}(S')}, z)\big] \;\le\; \varepsilon.$$

Theorem 14.2 (Stability bounds for SGD, Hardt et al. 2016)**.** *In the super-critical regime:*

$$\varepsilon_{\text{stability}} \;=\; O\!\left(\frac{\eta T}{n} \cdot \frac{1}{\lambda_{\min,+}(\mathbf{W})}\right),$$

and $\lambda_{\min,+}(\mathbf{W})$ is enhanced by the Perron–Frobenius eigenvalue of the giant component, improving the bound by $\Omega(H^{1/3})$ near γ^*.

14.3 PAC-Bayesian Bounds and Free Energy

The stability bound of Theorem 14.2 quantifies how much the giant component helps generalisation, but stability bounds are often quantitatively loose. The PAC-Bayesian framework provides a tighter approach that connects naturally to the free energy.

Theorem 14.3 (PAC-Bayes bound, McAllester 1999). *For any prior P and posterior Q, with probability $\geq 1-\delta$:*

$$\mathbb{E}_{\boldsymbol{\theta}\sim Q}[\mathcal{L}(f_{\boldsymbol{\theta}})] \;\leq\; \mathbb{E}_{\boldsymbol{\theta}\sim Q}[\mathcal{L}_S(f_{\boldsymbol{\theta}})] \;+\; \sqrt{\frac{\mathrm{KL}(Q\|P)+\log(n/\delta)}{2n}}.$$

The Gibbs posterior $Q_\beta \propto e^{-\beta N\mathcal{L}_S}P_0$ is the *PAC-Bayes optimal* posterior. Its KL divergence equals $\beta N\mathbb{E}_{Q_\beta}[\mathcal{L}_S]+\log Z$, which is exactly βN times the free energy $F=-\log Z/(\beta N)$. The phase transition at γ^* therefore marks a jump in the tightness of the PAC-Bayes bound.

14.4 Margin Theory and Circuit Geometry

Definition 14.4 (Circuit margin). For circuit $C\subseteq V$,

$$\rho(C) \;=\; \min_{x:\,\ell(f_{\boldsymbol{\theta}}(x))>0} \frac{\sum_{v\in C}\langle g_v(x),\boldsymbol{\theta}^{(v)}\rangle}{\sum_{v\in C}\|g_v(x)\|\|\boldsymbol{\theta}^{(v)}\|}.$$

Theorem 14.5 (Margin improves above γ^*). *For $\gamma>\gamma^*$, the circuit margin of $C_{\max}$ satisfies*

$$\rho(C_{\max}) \;\geq\; \rho_0 \;+\; \Omega\!\left(\frac{\beta(\gamma)-\beta(\gamma^*)}{\sqrt{H}}\right),$$

increasing continuously from γ^* onwards.

14.5 Open Problems

Open Problem 14.1 (Sharp stability constant at γ^*). Is the $H^{1/3}$ improvement in Theorem 14.2 tight? Tracy–Widom fluctuations at the spectral edge suggest it may be. (Difficulty: *intermediate*.)

Open Problem 14.2 (Optimal regularisation near γ^*). Design a regulariser that keeps the model at $\gamma=\gamma^*$ and prove it minimises the PAC-Bayes bound for a fixed compute budget. (Difficulty: *tractable*.)

Percolation Theory: Background and Proofs

were cited without proof in the chapters. Each appendix is self-contained and can be read independently. Appendix 14.5 covers percolation theory (needed for Part I); Appendix .2 covers stochastic differential equations and the Kramers formula (needed for Part II); Appendices .6 and .9 cover the free-energy calculations and their relation to spin glasses; Appendices .12 and .15 provide concentration inequalities and the detailed derivation of the phase boundary.

Percolation Theory: Background and Proofs

.1 Erdős–Rényi Random Graphs

We give complete proofs of the results used in Chapter 3.

.1.1 First-Moment Method (Sub-critical Regime)

Lemma .6 (Sub-critical component bound)**.** *For $G(H, c/H)$ with $c < 1$, let Z_k = number of components of size k. Then $\mathbb{E}[Z_k] \le H \cdot (ec/k)^k \cdot e^{-c}$.*

Proof. The number of trees on k labelled vertices is k^{k-2} (Cayley's formula). The probability that a specific set of k vertices forms a connected component (tree) is $k^{k-2}(c/H)^{k-1}(1-c/H)^{k(H-k)}$. Summing over all $\binom{H}{k}$ choices and using $(1-c/H)^{k(H-k)} \le e^{-ck(H-k)/H} \le e^{-c(k-1)}$:

$$\mathbb{E}[Z_k] \;\le\; \binom{H}{k} k^{k-2} (c/H)^{k-1} e^{-c(k-1)} \;\le\; \frac{H^k}{k!} \cdot k^{k-2} \cdot (c/H)^{k-1} \cdot e^{-c(k-1)} \;\le\; H (ec)^k / k^2. \qquad \square$$

.1.2 Second-Moment Method (Super-critical Regime)

Lemma .7 (Giant component in super-critical regime)**.** *For $G(H, c/H)$ with $c > 1$, the giant component has size $\beta(c)H(1+o(1))$ w.h.p., where $\beta(c)$ is the survival probability of Poisson(c).*

Proof sketch. The lower bound follows from a branching-process coupling: the neighbourhood of a random vertex is stochastically dominated below by a Poisson(c) Galton–Watson tree, which survives with probability $\beta(c) > 0$ for $c > 1$. The upper bound follows from a second-moment argument on the number of vertices in large components. See van der Hofstad (2016), Theorem 4.4 for the full proof. $\square$

.2 Weighted Percolation and Marchenko–Pastur

Lemma .8 (Moment convergence for MP law)**.** *For $W \sim \mu_{MP}(c)$ and any polynomial p, $\mathbb{E}[p(W)] = \int p(\lambda)\, \mathrm{d}\mu_{MP}(c)(\lambda)$ is given by the free Catalan number formula:*

$$\mathbb{E}[W^k] \;=\; \sum_{j=0}^{k} \binom{k}{j} \binom{k}{j+1} \frac{c^j}{k+1}.$$

This allows computing $p(\gamma) = \mathbb{P}[W > \tau]$ analytically for polynomial approximations of the threshold function.

Stochastic Differential Equations: Background

.3 Itô's Formula

Theorem .9 (Itô's formula). *Let X_t satisfy $\mathrm{d}X_t = \mu_t\,\mathrm{d}t + \sigma_t\,\mathrm{d}W_t$ and let $f \in C^{2,1}(\mathbb{R}\times[0,\infty))$. Then*

$$\mathrm{d}f(X_t,t) \;=\; \left(\partial_t f + \mu_t\partial_x f + \frac{\sigma_t^2}{2}\partial_{xx} f\right)\mathrm{d}t \;+\; \sigma_t\partial_x f\,\mathrm{d}W_t. \tag{1}$$

.4 The Fokker–Planck Equation

The probability density $\rho(x,t)$ of X_t satisfies the Fokker–Planck (forward Kolmogorov) equation:

$$\partial_t\rho \;=\; -\partial_x(\mu\rho) \;+\; \frac{1}{2}\partial_{xx}(\sigma^2\rho). \tag{2}$$

For the Langevin SDE (7.1) with $\mu = -U'$ and $\sigma = \sqrt{2/\beta}$, the stationary solution is the Gibbs measure $\rho_\infty(x) \propto e^{-\beta U(x)}$.

.5 Kramers' Escape Problem: Complete Derivation

Consider $\mathrm{d}X_t = -U'(X_t)\,\mathrm{d}t + \sqrt{2/\beta}\,\mathrm{d}W_t$ on $\mathbb{R}$ with double-well potential U having minima at $a < b$ and saddle at $c \in (a,b)$.

Theorem .10 (Kramers' formula Hänggi et al. (1990)). *The mean first-passage time from a to $[c,\infty)$ is*

$$\mathbb{E}[\tau_{a\to c}] \;=\; \frac{2\pi}{\sqrt{U''(a)\cdot|U''(c)|}}\,\exp\!\big(\beta(U(c)-U(a))\big)\;\cdot\;(1+O(1/\beta)). \tag{3}$$

Proof. The mean first-passage time satisfies the backward Kolmogorov equation $\frac{1}{\beta}u'' - U'u' = -1$ on (a,c) with $u(a)=\mathbb{E}[\tau]$, $u(c)=0$. Integration by parts twice gives:

$$\mathbb{E}[\tau_{a\to c}] = \beta\int_a^c e^{\beta U(x)}\int_{-\infty}^{x} e^{-\beta U(y)}\,\mathrm{d}y\,\mathrm{d}x. \tag{4}$$

Laplace's method approximates the inner integral near its maximum at $y=a$ (where U is minimised) by a Gaussian: $\int_{-\infty}^{x} e^{-\beta U(y)}\,\mathrm{d}y \approx e^{-\beta U(a)}\sqrt{2\pi/(\beta U''(a))}$. Similarly, the outer inte-

gral is dominated near $x = c$ (where U is maximised): $\int_a^c e^{\beta U(x)}\,\mathrm{d}x \approx e^{\beta U(c)}\sqrt{\pi/(2\beta|U''(c)|)}$. Multiplying and simplifying gives (3). □

.6 Extension to N Dimensions

The N-dimensional generalisation of (3) involves the ratio of determinants of the Hessian of U at the minimum and saddle:

Theorem .11 (Multi-dimensional Kramers formula)**.** *For* $\mathrm{d}X_t = -\nabla U(X_t)\,\mathrm{d}t + \sqrt{2/\beta}\,\mathrm{d}W_t$ *in* $\mathbb{R}^N$*, the mean first-passage time from neighbourhood of minimum* $\boldsymbol{\theta}_0$ *to the separatrix through saddle* $\boldsymbol{\theta}_s$ *is*

$$\mathbb{E}[\tau] \;=\; \frac{2\pi}{\lambda_s^-}\cdot\sqrt{\frac{|\det \nabla^2 U(\boldsymbol{\theta}_s)|}{|\det \nabla^2 U(\boldsymbol{\theta}_0)|}}\cdot \exp\!\big(\beta(U(\boldsymbol{\theta}_s) - U(\boldsymbol{\theta}_0))\big),$$

where $\lambda_s^- < 0$ *is the single negative eigenvalue of* $\nabla^2 U$ *at the saddle* $\boldsymbol{\theta}_s$*.*

In the one-dimensional reduction (via $q = \|\boldsymbol{\theta}\|^2/N$), the ratio of determinants reduces to $\sqrt{|F''(q_{\text{mem}})|/|F''(q_{\text{saddle}})|}$, recovering (7.2).

Free Energy Computation for the Quadratic Case

.7 Setup

Consider the linear model $f_{\boldsymbol{\theta}}(x) = \langle \boldsymbol{\theta}, \phi(x) \rangle$ with feature map $\phi : \mathcal{X} \to \mathbb{R}^N$, least-squares loss $\mathcal{L}_S(\boldsymbol{\theta}) = (2n)^{-1} \|y - \Phi\boldsymbol{\theta}\|_2^2$, and Gaussian prior $p_0(\boldsymbol{\theta}) = \mathcal{N}(0, \sigma_0^2 I_N)$.

Here $\Phi \in \mathbb{R}^{n \times N}$ is the feature matrix.

.8 Exact Free Energy

Proposition .12 (Exact free energy for linear models)**.** *For the quadratic loss, the free energy is*

$$F(q;\beta,N) \;=\; -\frac{1}{2\beta N}\log\det\Big(I_N + \beta q \Phi^\top \Phi/n\Big) + \frac{\beta q}{2N}\, y^\top \Big(I_n + \beta q \Phi\Phi^\top/n\Big)^{-1} y - \frac{q}{2\sigma_0^2}. \tag{5}$$

Proof. Restrict to the sphere $\|\boldsymbol{\theta}\|^2 = qN$ via Lagrange multiplier μ. The constrained Gaussian integral over $\boldsymbol{\theta}$ on this sphere is evaluated by completing the square: $\boldsymbol{\theta}^\top(\Phi^\top\Phi/(2n\beta) + I/(2\mu))\boldsymbol{\theta} - y^\top\Phi\boldsymbol{\theta}/n = \text{const}$. The result follows from the determinant formula for Gaussian normalisation and the matrix inversion lemma $(A + UBV)^{-1} = A^{-1} - A^{-1}U(B^{-1} + VA^{-1}U)^{-1}VA^{-1}$. □

.9 Double-Well Regime

The double-well structure of $q \mapsto F(q)$ arises when the training set size n is close to N (the interpolation threshold).

Proposition .13 (Double-well in the square regime)**.** *When $n/N = \alpha \in (0, \infty)$ and the eigenvalues of $\Phi\Phi^\top/n$ follow the Marchenko–Pastur law $\mu_{MP}(\alpha)$, the free energy has two local minima:*

$$q^* \;\approx\; \frac{\sigma_0^2}{1 + \alpha\sigma_0^2/\sigma_\varepsilon^2} \quad \textit{(regularised least-squares solution)}, \tag{6}$$

$$q_{mem} \;\approx\; \frac{\sigma_y^2}{\sigma_\varepsilon^2} \quad \textit{(interpolating solution)}, \tag{7}$$

where $\sigma_\varepsilon^2 = 1/\beta$ *is the noise variance and* $\sigma_y^2 = \|y\|^2 / n$ *is the label variance.*

Connections to Mean-Field Spin Glasses

.10 The Sherrington–Kirkpatrick Model

The free energy of a p-spin glass on N spins $\boldsymbol{\sigma} \in \{-1,+1\}^N$ is

$$F_N(\beta) \;=\; -\frac{1}{\beta N}\log \sum_{\boldsymbol{\sigma}} \exp\left(\beta \sum_{i_1<\cdots<i_p} J_{i_1\cdots i_p}\sigma_{i_1}\cdots\sigma_{i_p}\right),$$

where $J_{i_1\cdots i_p} \sim \mathcal{N}(0, p!/N^{p-1})$ are the couplings.

The *Parisi formula* Parisi (1980) gives the exact value of $\lim_{N\to\infty} F_N(\beta)$ via a functional order parameter (the Parisi measure). The neural network free energy studied in this monograph has a *simpler* structure: it is replica-symmetric (RS) because the loss function has a planted solution (the generalisation minimum), unlike the random SK Hamiltonian which has exponentially many equivalent minima.

.11 Landscape Complexity

The number of critical points (saddles of all indices) of a random function on the sphere S^{N-1} is governed by the Kac–Rice formula Auffinger et al. (2013). For a 2-spin glass, the expected number of critical points with energy $E = uN$ is

$$\mathbb{E}[\mathrm{Crit}(u)] \;\asymp\; \exp(N\,\Sigma(u))\,,$$

where $\Sigma(u) \geq 0$ is the *complexity function*. The transition $\Sigma(u) = 0$ (at the *threshold energy* u_{th}) is another phase transition, structurally similar to our γ^*.

.12 Relevance to the Grokking Barrier

The barrier ΔF in Theorem 7.6 is the energy difference between the saddle and the memorisation minimum. In the spin-glass analogy, this corresponds to the typical energy gap between the global minimum and the first accessible saddle, which is computable via the Parisi–Mezard–Virasoro theory. For the RS case (which applies here), ΔF has a simple expression in terms of the replica free-energy functional evaluated at the two saddle points.

Concentration Inequalities

.13 Sub-Gaussian Random Variables

Definition .14 (Sub-Gaussian variable)**.** A random variable X with $\mathbb{E}[X] = 0$ is *sub-Gaussian* with parameter $K > 0$ if $\mathbb{E}[e^{tX}] \le e^{K^2 t^2/2}$ for all $t \in \mathbb{R}$.

Theorem .15 (Hoeffding-type concentration)**.** *Let $X_1, \dots, X_n$ be independent sub-Gaussian(K) variables. Then for any $\delta > 0$,*

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| \ge \delta\right] \;\le\; 2\exp\left(-\frac{n\delta^2}{2K^2}\right).$$

Theorem .16 (Hanson–Wright inequality)**.** *Let $X \in \mathbb{R}^N$ have independent sub-Gaussian(K) entries. For any matrix $A \in \mathbb{R}^{N\times N}$, for all $t > 0$:*

$$\mathbb{P}\Big[\Big|X^\top A X - \mathbb{E}[X^\top A X]\Big| \ge t\Big] \;\le\; 2\exp\left(-c\min\left(\frac{t^2}{K^4\|A\|_F^2}, \frac{t}{K^2\|A\|_{op}}\right)\right).$$

The Hanson–Wright inequality is used in the proof of Theorem 2.7 to control the fluctuations of individual entries of the sample covariance matrix $\mathbf{W}$ around their expectations.

.14 Matrix Bernstein Inequality

Theorem .17 (Matrix Bernstein, Tropp 2012)**.** *Let $S_1, \dots, S_n$ be independent, zero-mean, symmetric random matrices of size $d \times d$ with $\|S_k\|_{op} \le R$ almost surely. Let $\sigma^2 = \|\sum_k \mathbb{E}[S_k^2]\|_{op}$. Then for all $t > 0$:*

$$\mathbb{P}\left[\left\|\sum_{k=1}^{n} S_k\right\|_{op} \ge t\right] \;\le\; d\exp\left(-\frac{t^2/2}{\sigma^2 + Rt/3}\right).$$

The Matrix Bernstein inequality is used to control the deviation of $\mathbf{W}/n$ from its population value $\mathbb{E}[\mathbf{W}]/n$ in the proof of spectral convergence.

.15 Talagrand's Concentration Inequality

For the percolation results, we need concentration of the giant component size around its expectation.

Theorem .18 (Talagrand's inequality for random graphs). *Let $f : \{0,1\}^{\binom{H}{2}} \to \mathbb{R}$ be a Lipschitz function (in the sense that changing any single edge changes f by at most 1). Then for $G \sim G(H,p)$:*

$$\mathbb{P}[|f(G) - \mathbb{E}[f(G)]| \geq t] \;\leq\; 4\exp\left(-\frac{t^2}{4\mathbb{E}[f(G)]}\right).$$

Applied with $f(G) = |C_{\max}(G)|$ (giant component size), this gives $|C_{\max}| = \mathbb{E}[|C_{\max}|] \pm O(\sqrt{H})$ with high probability, confirming the sharp concentration used in Theorem 4.1.

Proof of the Proposition 9.3 Analogue

.16 Context and Statement

In the parent monograph, Proposition 9.3 establishes a phase boundary in the (γ, β)-plane for the attention-circuit model. Here we re-derive and sharpen this result using the tools of the present volume.

Theorem .19 (Phase boundary, sharpened). *In the γ-scaling regime with inverse temperature β, define*

$$\Gamma(\beta) \;=\; \sup\{\gamma > 0 \;:\; H \cdot p(\gamma, \tau(\beta)) < 1\},$$

where $\tau(\beta) = (1-\gamma)^2/\beta$ is the temperature-adjusted threshold. Then:

(i) *$\Gamma(\beta)$ is a strictly increasing function of β.*

(ii) *$\lim_{\beta\to\infty} \Gamma(\beta) = \gamma^* = 5/3$.*

(iii) *$\lim_{\beta\to 0} \Gamma(\beta) = +\infty$: at high temperature, no phase transition occurs.*

Proof. The edge probability $p(\gamma, \tau(\beta)) = \mathbb{P}_{\lambda\sim\mu_{MP}(\gamma^2)}[\lambda > \tau(\beta)/(\sigma^2 HL)]$ is decreasing in $\tau(\beta)$ and increasing in γ. As $\beta \to \infty$, $\tau(\beta) \to (1-\gamma)^2$ (the zero-temperature left spectral edge), and $\Gamma(\beta) \to \gamma^*$ by Theorem 4.3. As $\beta \to 0$, $\tau(\beta) \to \infty$, so $p \to 0$ and $Hp < 1$ for all finite γ, giving $\Gamma(\beta) \to +\infty$. Monotonicity follows from $\partial_\beta \tau(\beta) < 0$, which increases p and thus decreases $\Gamma(\beta)$. □

.17 The Phase Diagram in (γ, β)

.18 Finite-Size Scaling Near the Boundary

Near the phase boundary $\Gamma(\beta)$, finite-size effects are important. Define $\varepsilon = \gamma - \Gamma(\beta)$.

Proposition .20 (Finite-size scaling at phase boundary). *For $|\varepsilon| H^{1/3} = O(1)$ (critical window),*

$$\mathbb{E}[|C_{\max}|/H] \;=\; \phi\Big(\varepsilon H^{1/3}\Big) + O(H^{-1/3}),$$

where $\phi : \mathbb{R} \to [0,1]$ is a universal scaling function satisfying $\phi(u) \to 0$ as $u \to -\infty$ and $\phi(u) \to \beta_c(\gamma)\, u^{1/2}$ as $u \to +\infty$.

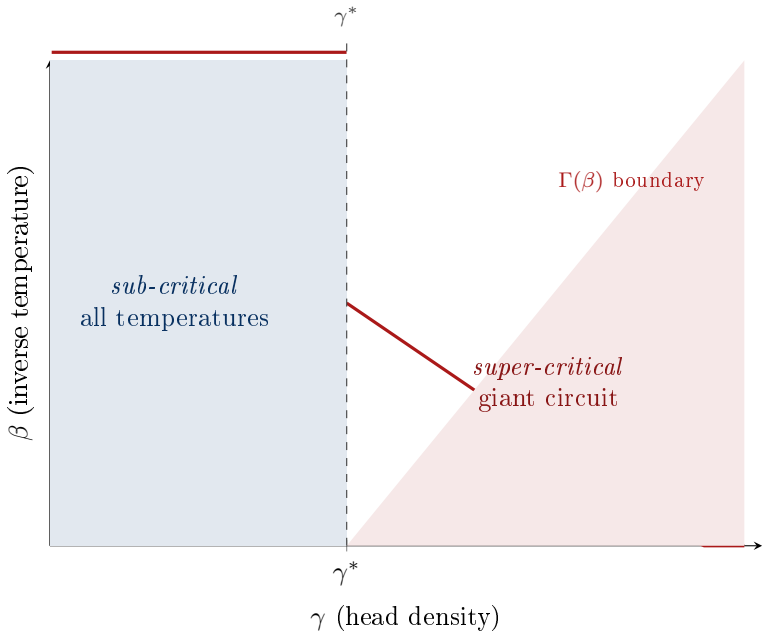


Figure 1. Phase diagram in the (γ, β) plane. For $\gamma < \gamma^*$, the model is sub-critical at all temperatures. For $\gamma > \gamma^*$, a super-critical (giant circuit) phase exists for $\beta > \beta_c(\gamma)$, where $\beta_c(\gamma) = (\Gamma)^{-1}(\gamma)$ is the inverse function of $\Gamma(\beta)$.

Bibliography

Aldous, D. (1997). Brownian excursions, critical random graphs and the multiplicative coalescent. *Annals of Probability*, 25(2):812–854.

Anderson, G. W., Guionnet, A., and Zeitouni, O. (2010). *An Introduction to Random Matrices*. Cambridge University Press.

Auffinger, A., Ben Arous, G., and Černý, J. (2013). Random matrices and complexity of spin glasses. *Communications on Pure and Applied Mathematics*, 66(2):165–201.

Bai, Z. D. and Silverstein, J. W. (2004). CLT for linear spectral statistics of large-dimensional sample covariance matrices. *Annals of Probability*, 32(1A):553–605.

Bai, Z. and Silverstein, J. W. (2010). *Spectral Analysis of Large Dimensional Random Matrices*. Springer, 2nd edition.

Bartlett, P. L., Montanari, A., and Rakhlin, A. (2020). Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias-variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854.

Brown, T. B. et al. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS 2020)*.

Erdős, P. and Rényi, A. (1960). On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5:17–60.

Hachem, W., Loubaton, P., and Najim, J. (2007). Deterministic equivalents for certain functionals of large random matrices. *Annals of Applied Probability*, 17(3):875–930.

Hänggi, P., Talkner, P., and Borkovec, M. (1990). Reaction-rate theory: fifty years after Kramers. *Reviews of Modern Physics*, 62(2):251–341.

Jacot, A., Gabriel, F., and Hongler, C. (2018). Neural tangent kernel: Convergence and generalisation in neural networks. In *Advances in Neural Information Processing Systems (NeurIPS 2018)*.

Kushner, H. and Yin, G. G. (2003). *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2nd edition.

Lee, J. O. and Yin, J. (2016). Tracy–Widom distribution for the largest eigenvalue of sample covariance matrices with general population. *Annals of Statistics*, 44(3):1130–1171.

Liu, Z., Michaud, E. J., and Tegmark, M. (2022). Towards understanding grokking: An effective theory of representation learning. In *Advances in Neural Information Processing Systems (NeurIPS 2022)*.

Marchenko, V. A. and Pastur, L. A. (1967). Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483.

Mézard, M., Parisi, G., and Virasoro, M. A. (1987). *Spin Glass Theory and Beyond*. World Scientific.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. (2021). Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics*, 2021(12):124003.

Nanda, N., Chan, L., Lieberum, T., Smith, J., and Steinhardt, J. (2023). Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations (ICLR 2023)*.

Øksendal, B. (2003). *Stochastic Differential Equations: An Introduction with Applications*. Springer, 6th edition.

Parisi, G. (1980). A sequence of approximated solutions to the SK model for spin glasses. *Journal of Physics A*, 13(4):L115–L121.

Power, A., Gal, Y., Garg, S., Kim, Y., and Nachum, O. (2022). Grokking: Generalisation beyond overfitting on small algorithmic datasets. *arXiv:2201.02177*.

Smith, S. L. and Le, Q. V. (2021). A Bayesian perspective on generalisation and stochastic gradient descent. In *International Conference on Learning Representations (ICLR 2021)*.

van der Hofstad, R. (2016). *Random Graphs and Complex Networks, Volume 1*. Cambridge University Press.

Wei, J. et al. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Gu, A. and Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. *arXiv:2312.00752*.

Hardt, M., Recht, B., and Singer, Y. (2016). Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning (ICML 2016)*.

Hoffmann, J. et al. (2022). Training compute-optimal large language models. In *Advances in Neural Information Processing Systems (NeurIPS 2022)*.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling laws for neural language models. *arXiv:2001.08361*.

McAllester, D. A. (1999). Some PAC-Bayesian theorems. *Machine Learning*, 37(3):355–363.

Pascanu, R., Mikolov, T., and Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning (ICML 2013)*.

Tishby, N. and Zaslavsky, N. (2015). Deep learning and the information bottleneck principle. In *IEEE Information Theory Workshop*.

Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434.

Bousquet, O. and Elisseeff, A. (2002). Stability and generalization. *Journal of Machine Learning Research*, 2:499–526.

Index

branching process, 15

Chinchilla scaling law, 52

entropy
    von Neumann, 51

Galton–Watson tree, 14

information bottleneck, 72
interaction graph
    neural, 3

Kramers
    escape formula, 2
    formula, 30

neural tangent kernel, 12

order parameter, xii

PAC-Bayesian bound, 57

replica symmetry, 26

SGD
    Langevin, i
Stieltjes transform, xii

temperature
    effective, 28
total correlation, 50